I ACT THEREFORE I AM: WHEN IS JEPA'S ACTION-CONDITIONING ENOUGH TO LEARN CAUSAL MECHANISMS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent empirical and theoretical advances suggest that joint-embedding predictive architectures (JEPAs) may learn meaningful representations for action-conditioned prediction of future outcomes, thus becoming one of the foundational structures for world models. However, accurate prediction does not, in general, necessarily imply recovery of underlying causal states that give rise to the observed dynamics. This work investigates when and how JEPAs can recover the underlying causal states from observations. We first introduce a latent variable model, in which high-dimensional observations are generated from latent causal states whose dynamics are governed by action-conditioned transition mechanisms. Based on this formulation, we develop a general information-theoretic objective that combines conditional likelihood maximization for learning transition dynamics with entropy maximization for preserving latent state information. We then establish identifiability conditions under which representations learned by this general objective recover the underlying latent causal states up to component-wise invertible transformations and permutation. One key condition for such identifiability is sufficient action-induced variation in the transition mechanisms. Guided by this finding, we instantiate the general objective with an action-modulated Gaussian additive-noise model, yielding action-modulated JEPA (A-JEPA). Experiments on synthetic environments verify the theoretical findings under the identifiability conditions and robustness to moderate violations, while visual benchmarks demonstrate improved state recovery and transfer to unseen transition mechanisms.

1 INTRODUCTION

World models seek to encode observations into compact, low-dimensional representations and learn action-conditioned transition dynamics among them, thereby capturing the consequences of possible actions and supporting planning and decision-making (Ha & Schmidhuber, 2018a; Hafner et al., 2019b; 2025). The joint-embedding predictive architecture (JEPA) provides a simple yet effective approach toward this goal by predicting future representations from past representations, conditioned on actions (LeCun et al., 2022). Empirically, this paradigm has achieved strong prediction performance (Assran et al., 2025; Maes et al., 2026; Destrade et al., 2025; Huang, 2026). Recent theoretical work has further provided formal guarantees for the predictive representations learned by JEPAs. In particular, under a specific Gaussian latent-variable model, JEPA can recover the underlying latent variables up to an orthogonal transformation, thereby supporting accurate future prediction (Klindt et al., 2026).

Despite recent progress in learning predictive representations, both empirically and theoretically, accurate prediction does not, in general, imply recovery of the underlying latent causal states governing the observed dynamics. In particular, representations may achieve accurate future prediction by exploiting statistical correlations without recovering the underlying causal mechanisms (Pearl, 2009; Schölkopf & von Kügelgen, 2023; Parascandolo et al., 2018; Schölkopf et al., 2021). Such correlations may fail to hold in previously unseen situations, such as under environmental changes, leading to degraded generalization (Ye et al., 2024; Schölkopf et al., 2021).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A natural way to address this limitation is to go beyond prediction alone and require world models to capture the underlying causal mechanisms governing the observed dynamics, giving rise to causal world models (Richens & Everitt, 2024; Zhu et al., 2022; Ceriscioli & Mohan, 2026). Recent work has begun to explore this direction within the JEPA framework (Nam et al., 2026). However, this approach relies primarily on heuristically imposed causal inductive biases, and it still urgently needs formal guarantees for recovering the underlying causal states from learned representations. This work therefore asks the following question:

> When and how can JEPA recover the underlying causal states from observational data?

To answer the question, this work makes the following contributions.

- We introduce a latent variable model to formulate JEPA in Sec. 2, in which observations are generated from latent causal states evolving through action-conditioned transition mechanisms. Based on this, we develop a general information-theoretic objective that combines conditional likelihood maximization with entropy maximization for learning predictive and information-preserving representations.
- Building on the formulation above, we establish an identifiability theory in Sec. 3, showing that, under certain assumptions, representations learned by the proposed general objective recover the latent causal states up to component-wise invertible transformations and permutation. *To the best of our knowledge, this is the first component-wise identifiability result within the JEPA framework.*
- Guided by the theoretical finding, we instantiate the general objective with an action-modulated Gaussian additive-noise model, yielding action-modulated JEPA (A-JEPA) in Sec. 4.
- We conduct extensive experiments in Sec. 5 on simulated data, verifying the theoretical findings when the identifiability assumptions hold and demonstrating robustness under moderate violations. Experiments on visual benchmarks further show substantially improved component-wise latent-state recovery and strong transferability to unseen transition mechanisms.

2 A Latent Variable Model and an Information-Theoretic JEPA

To study when and how JEPA can recover the underlying latent causal states, we first formulate the observed action-conditioned dynamics through a latent variable model. Under this formulation, we examine what a JEPA objective requires to support latent-state recovery, motivating a general information-theoretic objective that combines transition predictiveness with information preservation. We then characterize the global optima of this objective, providing the foundation for the identifiability analysis in Sec. 3.

2.1 Setup: A Latent Causal Generative Model

Consider a latent causal generative process indexed by time t, with latent state $\mathbf{s}_t \in \mathbb{R}^d$, observed action $\mathbf{a}_t$, and high-dimensional observation $\mathbf{x}_t$. We assume that the latent dynamics are governed by a temporal structural causal model (SCM) whose associated graph is a directed acyclic graph, such that each latent variable $s_{t,i}$ depends only on a subset of variables in the previous state. Let $\mathrm{pa}_i \subseteq \{1, \dots, d\}$ denote the indices of the causal parents of $s_{t,i}$ in $\mathbf{s}_{t-1}$. The action $\mathbf{a}_{t-1}$ is treated as an observed exogenous input that may modulate each transition mechanism.

We model the latent causal dynamics and observation process as:

$$s_{t,i} := f_i\left(\mathbf{s}_{t-1,\mathrm{pa}_i}, \mathbf{a}_{t-1}, \epsilon_{t,i}\right), \quad i = 1, \dots, d, \tag{1}$$

$$\mathbf{x}_t := \mathbf{g}(\mathbf{s}_t). \tag{2}$$

Here, f_i denotes the structural causal mechanism governing the transition of the i-th latent variable, $\epsilon_{t,i}$ is the corresponding exogenous disturbance, and $\mathbf{g}$ is an unknown deterministic observation mapping. Together, Eqs. (1) and (2) define the latent causal generative process underlying the observed action-conditioned dynamics.

We assume that the exogenous disturbances $\epsilon_{t,1}, \dots, \epsilon_{t,d}$ are mutually independent and independent of $(\mathbf{s}_{t-1}, \mathbf{a}_{t-1})$, following the standard independent-noise assumption in structural causal models (Pearl,

2009; Peters et al., 2017; Schölkopf et al., 2021). The temporal SCM in Eq. (1) therefore induces the action-conditioned transition factorization:

$$p(\mathbf{s}_t \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) = \prod_{i=1}^{d} p\left(s_{t,i} \mid \mathbf{s}_{t-1,\mathrm{pa}_i}, \mathbf{a}_{t-1}\right). \tag{3}$$

2.2 A General Information-Theoretic JEPA Objective

Given the latent causal generative model in Eqs. (1) and (2), we investigate what a JEPA objective needs to learn representations that recover the underlying latent causal states $\mathbf{s}_t$. Conventional JEPA objectives primarily encourage prediction of the future representation from the previous representation and action. Intuitively, beyond supporting future prediction, latent-state recovery additionally requires the learned representation to retain sufficient information about the underlying latent state. This motivates two complementary principles:

- *Transition Predictiveness.* The learned representation should support accurate prediction of the future representation from the previous representation and action.
- *Information Preservation.* The learned representation should avoid discarding information about the underlying latent state that is necessary for its recovery.

We instantiate these two principles through conditional likelihood maximization and entropy maximization, respectively. Let $\mathbf{z}_t = \mathbf{h}(\mathbf{x}_t)$ denote the representation produced by an encoder $\mathbf{h}$, and let $p_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1})$ denote a parameterized action-conditioned transition model. As a result, we arrive at the following general information-theoretic JEPA objective:

$$\mathcal{L}_\lambda(\mathbf{h}, \boldsymbol{\phi}) = -\mathbb{E}\left[\log p_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1})\right] - \lambda H(\mathbf{z}_t), \tag{4}$$

where $H(\mathbf{z}_t)$ denotes the differential entropy of the learned representation and λ controls the trade-off between transition predictiveness and information preservation. The first term encourages the learned transition model to accurately characterize the action-conditioned prediction in representation space, while the second discourages the learned representation from discarding information about the underlying latent state by favoring high-entropy representations.

We next provide an initial theoretical analysis of the objective in Eq. (4) to examine the properties it provides toward latent state recovery. For a given encoder $\mathbf{h}$, the latent generative model together with the encoder induces an action-conditioned transition distribution in the representation space, which we denote by $Q_{\mathbf{h}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1})$. In contrast, $p_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1})$ denotes the parameterized transition model optimized to approximate this induced distribution. With these definitions, we obtain the following result.

Theorem 2.1. *Suppose the observations are generated according to the latent causal generative model in Eqs. (1)-(2). Let $\mathbf{z}_t = \mathbf{h}(\mathbf{x}_t)$ with $\mathbf{h} : \mathcal{X} \to (0,1)^d$, and let $(\mathbf{h}^\star, \boldsymbol{\phi}^\star)$ be a global minimizer of Eq. (4) with $\lambda \geq 1$ that attains its lower bound. Then*

$$I(\mathbf{z}_t; \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = I(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}), \tag{5}$$

$$p_{\boldsymbol{\phi}^\star}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = Q_{\mathbf{h}^\star}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) \quad \text{a.e.} \tag{6}$$

Proof. See Appendix C. □

Potential for Latent-State Recovery. Theorem 2.1 provides two key characterizations of the objective in Eq. (4) at a global optimum. First, the learned representation preserves all transition-predictive information contained in the underlying latent state, as characterized by Eq. (5). This ensures that the encoder $\boldsymbol{h}$ retains the information in the latent state that is relevant for predicting the transition dynamics. Second, the learned transition model matches the transition distribution induced in the representation space, as shown by Eq. (6). This ensures that $\boldsymbol{\phi}$ reproduces the induced action-conditioned transition distribution in the learned representation space. Together, these two properties ensure that the learned representation retains the information required to describe the latent transition, while the learned transition model faithfully reproduces the corresponding conditional distribution in representation space.

However, these properties only characterize the learned representation at the level of transition-relevant information preservation and conditional distribution matching, and do not by themselves

guarantee component-wise recovery of the individual latent causal states. That is, the learned representation may still correspond to non-trivial mixtures of the underlying latent causal states, e.g., through an orthogonal transformation. We therefore study additional conditions under which such mixing can be ruled out in the next section.

3 Identifiability Analysis: Answering When

In this section, we study conditions under which the learned representation recovers the latent causal states up to component-wise invertible transformations and permutation, i.e., component-wise identifiability. Without additional assumptions, such identifiability may not in general be guaranteed. We introduce the following conditions on the latent causal generative model in Eqs. (1) and (2).

Assumption 3.1. The mapping $\mathbf{g}$ is smooth and invertible.

Assumption 3.1 is to ensure that the observation process Eq. 2 does not discard information about the latent state, so that the latent state can, in principle, be recovered from the observation. Intuitively, if the observation mapping is non-invertible, distinct latent states may generate the same observation, making exact latent state recovery impossible in general.

Assumption 3.2. For any $\mathbf{s}_t$, there exist $2d+1$ actions $\mathbf{a}_{t-1}^{(0)}, \ldots, \mathbf{a}_{t-1}^{(2d)}$ such that the matrix

$$\mathbf{L} = \left[\mathbf{w}(\mathbf{s}_t, \mathbf{s}_{t-1}, \mathbf{a}_{t-1}^{(1)}) - \mathbf{w}(\mathbf{s}_t, \mathbf{s}_{t-1}, \mathbf{a}_{t-1}^{(0)}), \ldots, \mathbf{w}(\mathbf{s}_t, \mathbf{s}_{t-1}, \mathbf{a}_{t-1}^{(2d)}) - \mathbf{w}(\mathbf{s}_t, \mathbf{s}_{t-1}, \mathbf{a}_{t-1}^{(0)})\right] \tag{7}$$

is invertible. Here, for each i, $q_i = \log p\big(s_{t,i} \mid \mathbf{s}_{t-1,\mathrm{pa}_i}, \mathbf{a}_{t-1}\big)$, and

$$\mathbf{w}(\mathbf{s}_t, \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) := \left(\frac{\partial q_1}{\partial s_{t,1}}, \ldots, \frac{\partial q_d}{\partial s_{t,d}}, \frac{\partial^2 q_1}{\partial s_{t,1}^2}, \ldots, \frac{\partial^2 q_d}{\partial s_{t,d}^2}\right)^\top. \tag{8}$$

Intuitively, Assumption 3.2 requires different actions to induce sufficiently diverse changes in the latent transition mechanisms. Such action-induced variation provides distinct signatures for individual latent causal variables, allowing them to be distinguished from arbitrary mixtures.

With the main assumptions above, we can now establish the main identifiability result, as follows:

Theorem 3.3. *Suppose the latent causal states $\mathbf{s}_t$ and observations $\mathbf{x}_t$ follow the generative model defined in Eqs. (1) and (2). Assume that Assumptions 3.1 and-3.2 hold. Let $\mathbf{h} : \mathcal{X} \to (0,1)^d$ be any smooth encoder associated with a global minimizer of Eq. (4) with $\lambda > 1$. Then the true latent causal states $\mathbf{s}_t$ are related to the learned representations $\mathbf{z}_t$, which are obtained by the encoder $\mathbf{h}$, by the following relationship:*

$$\mathbf{z}_t = \big(r_1(s_{t,\pi(1)}), \ldots, r_d(s_{t,\pi(d)})\big)\,, \tag{9}$$

where π is a permutation and $r_1, \ldots, r_d$ are one-dimensional invertible transformations.

Proof. See Appendix D. □

Implication: Answering When. Theorem 3.3 shows that JEPA can recover the latent causal states up to component-wise invertible transformations and permutation. A key condition underlying this result is sufficient action variability, i.e., Assumption 3.2: different actions must induce sufficiently diverse changes in the latent transition mechanisms to rule out arbitrary mixing of the latent causal variables. This provides a primary answer to the "when" part of our central question: JEPA can identify the underlying latent causal states when the observed actions induce sufficiently rich variation in the transition mechanisms. Intuitively, if the transition mechanisms remain unchanged across actions, different latent mixtures may remain observationally indistinguishable, making component-wise recovery difficult.

4 From Identifiability to Practice: Answering How

Theorem 3.3 identifies sufficient action-induced variation in the latent transition mechanisms as a key condition for component-wise recovery. While the result applies to general nonparametric transition

mechanisms, it does not directly specify how such variation should be modeled and optimized in practice. We therefore consider a concrete transition model in which action-induced mechanism variation admits an explicit and tractable characterization.

To this end, we consider an action-modulated additive-noise model (ANM) with Gaussian noise as a representative instantiation. This choice is motivated by three considerations. First, ANMs provide a widely used class for modeling causal mechanisms (Hoyer et al., 2008; Peters et al., 2014; Bühlmann et al., 2014). Second, the Gaussian formulation yields a tractable conditional likelihood. Most importantly, it makes the action-variability condition in Assumption 3.2 directly interpretable through action-dependent conditional means and variances, as discussed below.

Action-Modulated Gaussian ANM in Latent Space. We instantiate the latent causal transition model in Eq. (3) as an action-modulated Gaussian ANM, as follows:

$$s_{t,i} = f_i(\mathbf{s}_{t-1,\mathrm{pa}_i}, \mathbf{a}_{t-1}) + \sigma_i(\mathbf{a}_{t-1})\epsilon_{t,i}, \qquad \epsilon_{t,i} \sim \mathcal{N}(0,1), \tag{10}$$

where f_i characterizes the deterministic transition mechanism and σ_i models the corresponding transition uncertainty. Both are allowed to vary with the action $\mathbf{a}_{t-1}$. Accordingly, the conditional transition distribution takes the form:

$$p(\mathbf{s}_t \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) = \prod_{i=1}^{d} \mathcal{N}\left(s_{t,i}; f_i(\mathbf{s}_{t-1,\mathrm{pa}_i}, \mathbf{a}_{t-1}), \sigma_i^2(\mathbf{a}_{t-1})\right). \tag{11}$$

Under this formulation, action-induced changes in the transition mechanisms are directly reflected by the changes in the conditional means f_i and variances σ_i^2. For this Gaussian model, the first- and second-order derivatives of the conditional log-density appearing in Assumption 3.2 take the form

$$\frac{\partial q_i}{\partial s_{t,i}} = -\frac{s_{t,i} - f_i}{\sigma_i^2}, \qquad \frac{\partial^2 q_i}{\partial s_{t,i}^2} = -\frac{1}{\sigma_i^2}, \tag{12}$$

where the dependence of f_i on $(\mathbf{s}_{t-1,\mathrm{pa}_i}, \mathbf{a}_{t-1})$ and of σ_i^2 on $\mathbf{a}_{t-1}$ is omitted for clarity. Eq. (12) provides a direct interpretation of Assumption 3.2. Action-induced changes in the conditional mean f_i affect the first-order derivative, whereas changes in the conditional variance σ_i^2 affect both the first- and second-order derivatives. Consequently, sufficient action variability requires actions to induce sufficiently rich variation in these conditional transition statistics so that the resulting matrix $\mathbf{L}$ has full rank.

Gaussian Transition Model in Representation Space. Following the Gaussian instantiation in the latent space, we parameterize the learned transition model to preserve the component-wise temporal causal structure. Specifically, let pa_i denote the parent set of the i-th representation component in the previous state. We model

$$p_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \prod_{i=1}^{d} \mathcal{N}\left(z_{t,i}; \mu_{\boldsymbol{\phi},i}(\mathbf{z}_{t-1,\mathrm{pa}_i}, \mathbf{a}_{t-1}), \sigma_{\boldsymbol{\phi},i}^2(\mathbf{a}_{t-1})\right). \tag{13}$$

Here, $\mu_{\boldsymbol{\phi},i}(\mathbf{z}_{t-1,\mathrm{pa}_i}, \mathbf{a}_{t-1})$ and $\sigma_{\boldsymbol{\phi},i}(\mathbf{a}_{t-1}) > 0$ denote the conditional mean and standard deviation of the i-th representation component, respectively. This parameterization mirrors the latent transition model by allowing each representation component to depend only on its temporal parents and the action. Implementations are detailed in Appendix F.

The corresponding negative log-likelihood (NLL) term in Eq. (4) is

$$\mathrm{NLL}_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \frac{1}{2}\sum_{i=1}^{d}\left[\frac{\left(z_{t,i} - \mu_{\boldsymbol{\phi},i}(\mathbf{z}_{t-1,\mathrm{pa}_i}, \mathbf{a}_{t-1})\right)^2}{\sigma_{\boldsymbol{\phi},i}^2(\mathbf{a}_{t-1})} + \log \sigma_{\boldsymbol{\phi},i}^2(\mathbf{a}_{t-1})\right]. \tag{14}$$

Here, we omit the additive constant $d\log(2\pi)/2$, which does not affect optimization.

The remaining challenge is to instantiate the entropy term in Eq. (4). Direct estimation of differential entropy in high-dimensional representation spaces is generally difficult. We therefore adopt a contrastive surrogate for the entropy term, motivated by the established asymptotic connection between contrastive learning with negative samples and entropy-regularized representation learning (Wang & Isola, 2020; Von Kügelgen et al., 2021). For a theoretical justification of this surrogate, see Sec. E.

Specifically, given an observed transition $(\mathbf{z}_{t-1}, \mathbf{a}_{t-1}, \mathbf{z}_t)$, we sample K alternative future representations $\mathbf{z}_t^{(1)}, \ldots, \mathbf{z}_t^{(K)}$ from other transitions in the minibatch and evaluate them under the same

action-conditioned transition model. Combining the NLL in Eq. (14) with this contrastive surrogate yields the practical objective

$$\mathcal{L} = \mathbb{E}\left[\alpha\,\mathrm{NLL}_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) + (1-\alpha)\log\frac{1}{K}\sum_{k=1}^{K}\exp\left(-\frac{\mathrm{NLL}_{\boldsymbol{\phi}}(\mathbf{z}_t^{(k)} \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1})}{\tau}\right)\right]. \tag{15}$$

Here, $\alpha \in (0,1)$ controls the trade-off between transition prediction and information preservation, while $\tau > 0$ denotes the temperature. The first term assigns high likelihood to the future representation, whereas the second discourages the same transition model from assigning high likelihood to alternative future representations. Together, the two terms encourage representations that are both transition-predictive and information-preserving, thus fulfilling our JEPA objective principles.

Implication: Answering How. Eq. (15) provides a practical instantiation of the two principles underlying Eq. (4). The Gaussian NLL implements transition predictiveness through conditional likelihood maximization, while the contrastive term provides a tractable surrogate for entropy-based information preservation. More importantly, the action-modulated Gaussian transition translates the abstract variability condition in Assumption 3.2 into a concrete modeling principle: by parameterizing action-dependent conditional means and variances, the learned transition model can capture the action-induced mechanism variation to ensure component-wise identifiability.

5 Empirical Evaluation

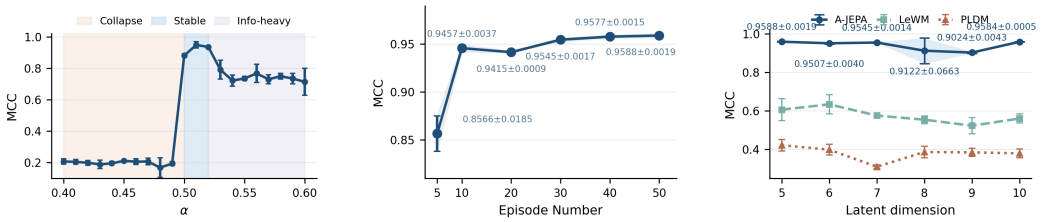
We first evaluate A-JEPA on synthetic systems generated according to the action-modulated latent transition model in Sec. 4, where the theoretical assumptions are satisfied by construction, to empirically examine the identifiability results developed above. We then consider controlled violations of these assumptions to assess how the method behaves when the theoretical conditions are only approximately satisfied. Finally, we evaluate on visual control environments to study whether the same design facilitates recovery of task-relevant state factors.

5.1 Simulation

Data generation. We generate synthetic temporal causal systems according to the action-modulated Gaussian ANM in Sec. 4, with ground-truth latent states and temporal DAGs. The setup provides controlled variation along the key dimensions of our analysis: latent dimensionality, action-induced mechanism variation, observation invertibility, and transition-noise specification. Actions modulate both the conditional means and variances of the latent transition mechanisms, while latent states are mapped to observations through an invertible nonlinear mixing function in the matched setting. This controlled setup allows us to evaluate component-wise state recovery, transition-structure recovery, and robustness to violations of the identifiability and modeling assumptions. See Sec. G for details.

Model and Evaluation. We train the encoder and action-conditioned transition model using the objective in Eq. (15). Consistent with the component-wise identifiability result in Theorem 3.3, we evaluate latent-state recovery using Spearman component-wise mean correlation coefficient (MCC). Specifically, we compute absolute Spearman correlations between learned and ground-truth latent coordinates, perform optimal matching using the Hungarian algorithm, and average the matched correlations. For transition-structure recovery, we align the learned coordinates using the same matching and evaluate the recovered temporal graph using edge F1, precision, recall, and structural Hamming distance (SHD).

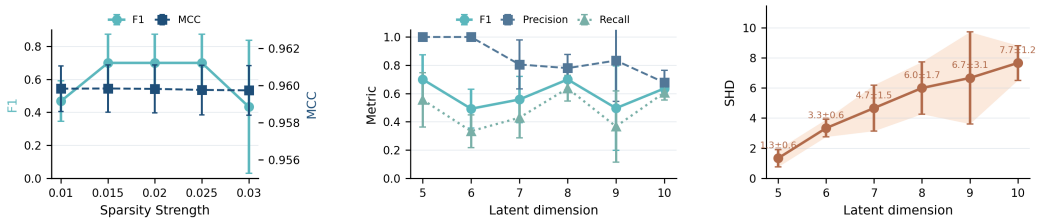
Latent-State Recovery Analysis. We first examine the factors governing latent-state recovery. Varying the trade-off parameter α in Eq. (15) reveals a sharp dependence on the balance between transition likelihood and information preservation (Fig. 1a). MCC remains low for $\alpha < 0.50$, peaks at 0.9479 near $\alpha = 0.51$, and then gradually decreases as greater weight is placed on transition likelihood. This highlights the importance of balancing transition predictiveness with information preservation for component-wise recovery. We next vary the number of distinct action settings. As shown in Fig. 1b, MCC improves from 0.8566 with five actions to 0.9588 with fifty, with gains gradually saturating. This trend is consistent with Assumption 3.2: richer action-induced mechanism variation provides stronger signals for distinguishing latent causal variables. Finally, Fig. 1c evaluates



(a) Trade-off hyperparameter α. (b) Action variability. (c) Latent dimension scaling.

Figure 1: Analysis of latent-state recovery. Left: effect of the trade-off between transition likelihood and information preservation. Middle: effect of increasing action-induced mechanism variation. Right: component-wise recovery across latent dimensions compared with JEPA-style baselines.

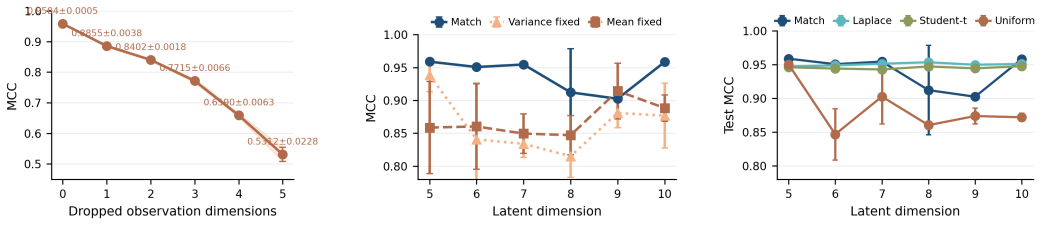
recovery as the latent dimension increases from $d = 5$ to $d = 10$. A-JEPA maintains MCC around 0.90-0.96, substantially outperforming LeWM-style and PLDM baselines. This shows that strong component-wise recovery persists as dimensionality increases.



(a) Parent-gate sparsity. (b) Graph recovery. (c) Structural Hamming distance.

Figure 2: Analysis of latent structure recovery. Left: effect of parent-gate sparsity on graph recovery. Middle: F1, precision, and recall across latent dimensions. Right: SHD across latent dimensions.

Transition-Structure Recovery Analysis. We further examine whether the recovered latent representations support identification of the underlying causal transition structure. As shown in Fig. 2a, latent-state recovery remains stable at MCC around 0.96, whereas graph recovery is more sensitive to the gate sparsity regularization (See Eq. (70) for implementation.), with edge F1 reaching approximately 0.70 in the best-performing range. We then evaluate graph recovery as the latent dimension increases from $d = 5$ to $d = 10$. Fig. 2b shows consistently high precision with moderate recall, while Fig. 2c shows increasing SHD as dimensionality grows. Together, these results indicate that the recovered representations retain meaningful information about the underlying transition structure, although graph recovery becomes increasingly challenging in higher dimensions.

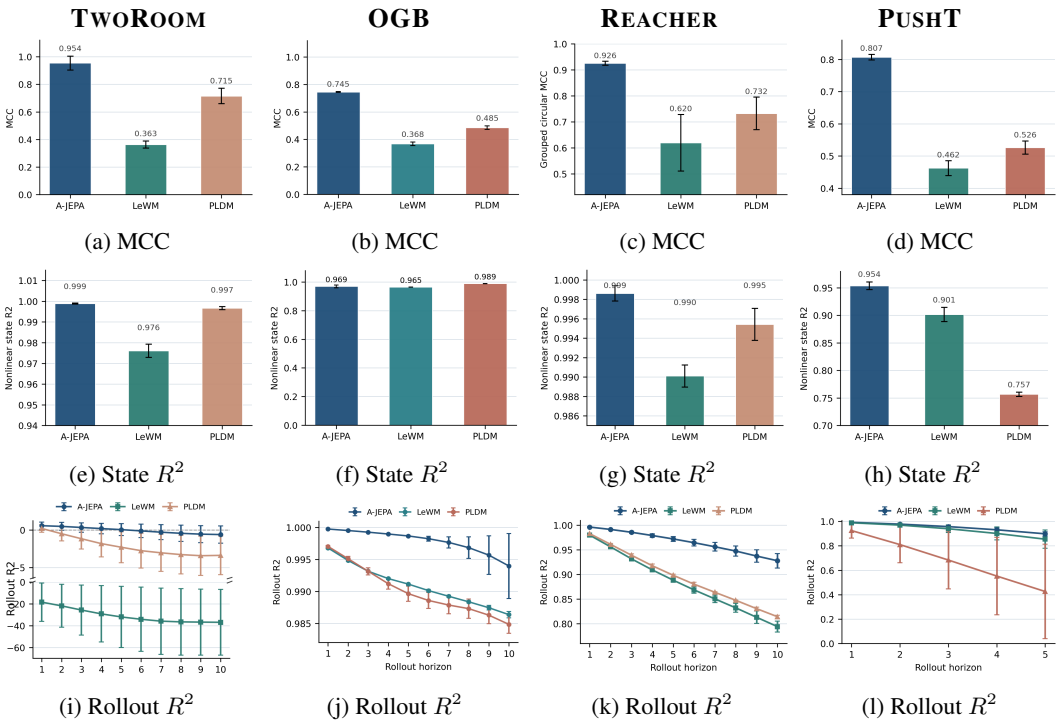


(a) Observation invertibility. (b) Action-induced variation. (c) Noise misspecification.

Figure 3: Assumption and model-specification analysis. Left: latent-state recovery degrades as the observation mapping becomes increasingly non-invertible. Middle: reducing action-induced variation in either the conditional mean or variance degrades recovery. Right: the proposed method remains robust to Laplace and Student-t noise, while uniform noise leads to a larger degradation.

Assumption and Model-Specification Analysis. Finally, we examine sensitivity to the assumptions and modeling choices underlying our framework. First, progressively removing observed

coordinates degrades MCC from 0.9584 to 0.5312 (Fig. 3a), consistent with the importance of an information-preserving observation mapping. Second, we weaken action-induced mechanism variation by allowing actions to modulate only the conditional mean or variance. Both restrictions degrade latent-state recovery across dimensions (Fig. 3b), supporting the role of rich action-induced variation in Assumption 3.2. Finally, replacing Gaussian noise with Laplace or Student-t noise causes little degradation, whereas uniform noise has a larger effect (Fig. 3c), indicating robustness to moderate transition-noise misspecification.



(a) MCC (b) MCC (c) MCC (d) MCC

(e) State R^2 (f) State R^2 (g) State R^2 (h) State R^2

(i) Rollout R^2 (j) Rollout R^2 (k) Rollout R^2 (l) Rollout R^2

Figure 4: Representation and transition recovery across four visual control environments.Columns correspond to TWOROOM, OGB, REACHER, and PUSHT, while rows report component-wise state recovery (MCC), joint-state information preservation (nonlinear state R^2), and multi-step action-conditioned prediction (rollout R^2), respectively. A-JEPA consistently improves component-wise recovery while retaining strong state information and predictive dynamics.

5.2 VISUAL CONTROL ENVIRONMENTS

We next examine whether the theoretical and empirical findings above extend beyond the controlled simulation setting to visual environments with more complex action-conditioned dynamics. We consider four environments: TWOROOM (Sobal et al., 2025), OGB (Park et al., 2025), REACHER (Tassa et al., 2018), and PUSHT (Zhou et al., 2025), covering increasingly challenging state transitions and interactions. We compare the proposed A-JEPA with two JEPA-style baselines, LeWM (Maes et al., 2026) and PLDM (Sobal et al., 2026). We evaluate the learned representations from three complementary perspectives. First, we measure component-wise recovery using MCC, which directly evaluates whether individual learned coordinates align with the underlying state factors. Second, we train a nonlinear decoder from the learned representations to the ground-truth states and report nonlinear state R^2, providing a proxy for how much joint-state information remains recoverable from the representation irrespective of coordinate-wise alignment. Third, we recursively apply the learned transition model under action sequences and report rollout R^2, measuring how well the learned representation supports action-conditioned transition prediction over increasing horizons. See Sec. H.

Component-Wise State Recovery. As shown in the first row of Fig. 4, A-JEPA consistently achieves higher MCC than the JEPA-style baselines across all four environments. The improvement is observed in both simpler environments such as TWOROOM and OGB, and more complex continuous-

control environments including REACHER and PUSHT. These results show that the component-wise recovery advantage persists in more complex visual environments.

Information Preservation Is Not Component-Wise Recovery. We next ask whether the lower MCC of the baselines merely reflects information loss. The middle row of Fig. 4 shows that all methods retain substantial information about the underlying joint state when evaluated with a nonlinear decoder. Importantly, high state information does not necessarily imply component-wise recovery. For example, on OGB, PLDM achieves nonlinear state $R^2 = 0.989$ but an MCC of only 0.485. This distinction is consistent with our theoretical analysis: information preservation alone does not rule out mixing between latent causal factors.

Action-Conditioned Transition Prediction. The bottom row of Fig. 4 evaluates recursive action-conditioned prediction over increasing rollout horizons. A-JEPA generally exhibits more stable long-horizon predictions, with particularly clear improvements on REACHER and PUSHT. Taken together, these results show that the improved component-wise recovery of A-JEPA is achieved while retaining substantial state information and strong predictive dynamics. We emphasize that this establishes an empirical association rather than implying that component-wise identifiability alone guarantees improved rollout performance.

5.3 ROBOT MECHANISM TRANSFER.

A key motivation for learning causal transition mechanisms is their potential to transfer beyond the environments in which they are learned. If the learned representation primarily captures environment-specific visual correlations, its transition dynamics are likely to degrade under environmental changes. In contrast, representations that capture causal action-conditioned mechanisms should support more stable predictive dynamics across such shifts. In our final experiment, we therefore examine whether the learned action-conditioned dynamics transfer beyond the training environments.

We construct a held-out-robot experiment on the RoboNet dataset (Dasari et al., 2019), which contains trajectories collected from multiple robot platforms. We train the models on Baxter, Kuka, and WidowX and evaluate them without adaptation on held-out Franka and Sawyer robots. This split introduces substantial changes in robot morphology, visual appearance, and embodiment. At test time, all model parameters are frozen. Given the encoded state of a held-out robot and its subsequent action sequence, we recursively roll the learned transition model forward for up to 20 steps. At each horizon h, we compare the predicted latent state $\hat{\mathbf{z}}_{t+h}$ with the encoding of the corresponding future observation, $\mathbf{z}_{t+h} = \mathbf{h}(\mathbf{x}_{t+h})$, using rollout R^2, where larger R^2 values denote more accurate latent rollout prediction. As shown in Fig. 5, A-JEPA maintains stable latent rollout predictions on both held-out robots, with rollout R^2 remaining above 0.90 at horizon 20 for both Franka and Sawyer. In contrast, matched-capacity LeWM and PLDM baselines exhibit substantially stronger degradation as the rollout horizon increases. Since all methods use the same encoder architecture and latent dimension and are evaluated without adaptation, the results provide evidence that A-JEPA learns action-conditioned latent dynamics with stronger transfer across unseen robot platforms.

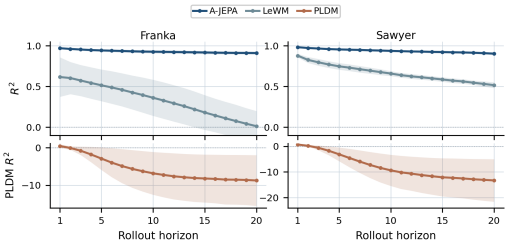


Figure 5: Held-out robot transfer on RoboNet. A-JEPA maintains substantially more stable latent rollout predictions on unseen Franka and Sawyer robots. See Sec. J for details.

6 CONCLUSION

We studied when and how JEPA can recover latent causal states from action-conditioned observations. We developed an information-theoretic objective and established component-wise identifiability under sufficient action-induced variation in the transition mechanisms, which motivates the practical A-JEPA formulation. Experiments on synthetic, visual, and held-out robot settings support the theory and show improved state recovery, robustness, and transfer of the learned dynamics.

AI USE STATEMENT

Generative AI tools were used to assist with language editing. All scientific ideas, theoretical results, experimental design, analysis, and conclusions were developed and verified by the authors. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

This work studies representation learning using publicly available models and datasets. It does not involve human subjects or private personal data.

REPRODUCIBILITY STATEMENT

We provide the assumptions and complete proofs of the theoretical results in the appendix. Detailed experimental protocols, dataset processing procedures, model configurations, and evaluation metrics are also provided in the appendix to facilitate reproduction of our results.

REFERENCES

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *ICML*, pp. 372–407. PMLR, 2023.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698*, 2023.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Sebastian Birk, Mohammad Vali Sanian, Amirhossein Vahidi, Samuel Ogden, Daniyal J Jafree, Adib Miraki Feriz, Carlo Leonardi, Arpit Merchant, Zichen He, Lloyd Steele, et al. Multi-scale modeling of human tissues from spatial transcriptomics with terra. *BioRxiv*, pp. 2026–07, 2026.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. *arXiv preprint arXiv:2203.16437*, 2022.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *arXiv preprint arXiv:2306.02235*, 2023.

Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Matteo Ceriscioli and Karthika Mohan. Agents robust to distribution shifts learn causal world models even under mediation. *Advances in Neural Information Processing Systems*, 38:77585–77619, 2026.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Lei Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. In *International Conference on Learning Representations*, volume 2026, pp. 89791–89807, 2026.

Guangyi Chen, Yifan Shen, Zhenhao Chen, Xiangchen Song, Yuewen Sun, Weiran Yao, Xiao Liu, and Kun Zhang. Caring: Learning temporal causal representation under non-invertible generation process. *arXiv preprint arXiv:2401.14535*, 2024.

Ciem Cornelissen, Sam Leroux, and Pieter Simoens. Le mumo jepa: Multi-modal self-supervised representation learning with learnable fusion tokens. *arXiv preprint arXiv:2603.24327*, 2026.

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019.

Fei Deng, Ingook Jang, and Sungjin Ahn. Dreamerpro: Reconstruction-free model-based reinforcement learning with prototypical representations. In *International conference on machine learning*, pp. 4956–4975. PMLR, 2022.

Matthieu Destrade, Oumayma Bounou, Quentin Le Lidec, Jean Ponce, and Yann LeCun. Value-guided action planning with jepa world models. *arXiv preprint arXiv:2601.00844*, 2025.

Seth Dobrin and Łukasz Chmiel. Identifiability without gaussianity: Symbolic world models and near-infinite temporal consistency. *arXiv preprint arXiv:2606.12471*, 2026.

Zhengcong Fei, Mingyuan Fan, and Junshi Huang. A-jepa: Joint-embedding predictive architecture can listen. *arXiv preprint arXiv:2311.15830*, 2023.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018a.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018b.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019b.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Nick Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, volume 2024, pp. 47376–47405, 2024.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*, 2022.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Naiwen Hu, Haozhe Cheng, Yifan Xie, Shiqi Li, and Jihua Zhu. 3d-jepa: A joint embedding predictive architecture for 3d self-supervised representation learning. *arXiv preprint arXiv:2409.15803*, 2024.

Yongchao Huang. Vjepa: Variational joint embedding predictive architectures as probabilistic world models. *arXiv preprint arXiv:2601.14354*, 2026.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *NeurIPS*, 29, 2016.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial intelligence and statistics*, pp. 460–469. PMLR, 2017.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *AISTAS*, pp. 2207–2217. PMLR, 2020.

Siwon Kim. A lightweight self-supervised learning framework for multivariate time series using hierarchical-jepa on ecg data. *arXiv preprint arXiv:2607.01145*, 2026.

David Klindt, Yann LeCun, and Randall Balestriero. When does lejepa learn a world model? *arXiv preprint arXiv:2605.26379*, 2026.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pp. 428–484. PMLR, 2022.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Hongyang Lei, Xiaolong Cheng, Qi Qin, Dan Wang, Kun Fan, Huazhen Huang, Qingqing Gu, Yetao Wu, Zhonglin Jiang, Yong Chen, et al. M3-jepa: Multimodal alignment via multi-gate moe based on the joint-embedding predictive architecture. *arXiv preprint arXiv:2409.05929*, 2024.

Guillaume Letellier, Antonin Vacheret, and Frédéric Jurie. Jetparticle-jepa: An efficient self-supervised representation learning method for jet tagging in high-energy physics. *arXiv preprint arXiv:2606.14813*, 2026.

Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Guangyi Chen, and Kun Zhang. On the identification of temporal causal representation with instantaneous dependence. In *International Conference on Learning Representations*, volume 2025, pp. 35203–35242, 2025.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, 17–23 Jul 2022.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent polynomial causal models through the lens of change. In *The Twelfth International Conference on Learning Representations*, 2024.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifying weight-variant latent causal models. *Journal of Machine Learning Research*, 27(1):1–49, 2026.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Naoki Morihira, Amal Nahar, Kartik Bharadwaj, Yasuhiro Kato, Akinobu Hayashi, and Tatsuya Harada. R2-dreamer: Redundancy-reduced world models without decoders or augmentation. *arXiv preprint arXiv:2603.18202*, 2026.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-jepa: Learning world models through object-level latent masking. *arXiv preprint arXiv:2602.11389*, 2026.

Tung D Nguyen, Rui Shu, Tuan Pham, Hung Bui, and Stefano Ermon. Temporal predictive coding for model-based planning in latent space. In *International conference on machine learning*, pp. 8130–8139. PMLR, 2021.

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pp. 4036–4044. PMLR, 2018.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations (ICLR)*, 2025.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014. URL http://jmlr.org/papers/v15/peters14a.html.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

Seth Rabinowitz, Prbhuv Nigam, Nicholas Santolla, and Colby T Ford. Kūkulu: Diffusion-based reconstruction of antibody cdr loops using a structure-aware joint embedding predictive architecture. *bioRxiv*, pp. 2026–08, 2026.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024.

Ayumu Saito, Prachi Kudeshia, and Jiju Poovvancheri. Point-jepa: A joint embedding predictive architecture for self-supervised learning on point cloud. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7348–7357. IEEE, 2025.

Bernhard Schölkopf and Julius von Kügelgen. From statistical to causal learning. In *International Congress of Mathematicians*, pp. 5540–5593. European Mathematical Society-EMS-Publishing House GmbH, 2023.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. *arXiv preprint arXiv:2007.05929*, 2020.

Anna Seigal, Chandler Squires, and Caroline Uhler. Linear causal disentanglement via interventions. *arXiv preprint arXiv:2211.16467*, 2022.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *The Journal of Machine Learning Research*, 23(1):10994–11048, 2022.

Ryan Shihabi, Siddhant Karmali, Brent Vaughan, Sharief Taraman, and Manolis Kellis. Hi-jepa: A world model of molecular organization learned from measured proximity. *bioRxiv*, pp. 2026–08, 2026.

Uladzislau Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim GJ Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. *Advances in Neural Information Processing Systems*, 38:43905–43941, 2026.

Vlad Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim GJ Rudner, and Yann LeCun. Stress-testing offline reward-free reinforcement learning: A case for planning with latent dynamics models. In *7th Robot Learning Workshop: Towards Robots with Human-Level Abilities*, 2025.

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. *Advances in Neural Information Processing Systems*, 36:8092–8113, 2023.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite, 2018.

Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 2023.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *NeurIPS*, 2021.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2023.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*, 2021.

Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, and Yang Gao. Mastering atari games with limited data. *Advances in neural information processing systems*, 34:25476–25488, 2021.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey. *arXiv e-prints*, pp. arXiv–2402, 2024.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

Haichao Zhang, Yijiang Li, Shwai He, Tushar Nagarajan, Mingfei Chen, Jianglin Lu, Ang Li, and Yun Fu. Thinkjepa: Empowering latent world models with large vision-language reasoning model. *arXiv preprint arXiv:2603.22281*, 2026a.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

Xiangteng Zhang, Yang Guan, Bo Zhang, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models. *arXiv preprint arXiv:2607.22430*, 2026b.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Forty-second International Conference on Machine Learning*, 2025.

Meng Zhou, Wenhao You, Yuxing Chen, and Yueying Tian. Beyond representation learning: A systematic study of joint-embedding predictive generation for 3d brain mri. *arXiv preprint arXiv:2608.28787*, 2026a.

Qi Zhou, Yuan Le, Xiaoning Qi, Hanshuo Lu, Shaole Chang, Yunjia Wu, Haoran Wang, Ruoyao Ran, and Xin Li. Cellos: Learning a world model of cellular state through joint embedding prediction. *bioRxiv*, pp. 2026–06, 2026b.

Zheng-Mao Zhu, Xiong-Hui Chen, Hong-Long Tian, Kun Zhang, and Yang Yu. Offline reinforcement learning with causal structured world models. *arXiv preprint arXiv:2206.01474*, 2022.

Appendix

Table of Contents

A Related Work 16

B Limitations 17

C Proof of Theorem 2.1 18

D Proof of Theorem 3.3 19

E Negative-Sample Contrastive Approximation to Entropy 22

F Implementation of A-JEPA 23

G Simulation Details 24

H Visual Benchmark Details 26

I Visual Intervention Experiments on REACHER 29

J Details of Visual Transfer Experiments 30

K Discussion I: LeWorldModel as a Constrained Instance 33

L Discussion II: Two Levels of Representation Collapse 34

A Related Work

Development of JEPA. The JEPA framework was originally inspired by the human brain's perception system (LeCun et al., 2022) to comprehend a dynamic environment in the representation space, which has become a core component of autonomous intelligent systems. Later, JEPA was employed as a self-supervised learning paradigm, such as shown in I-JEPA (Assran et al., 2023), MC-JEPA (Bardes et al., 2023), and LeJEPA (Balestriero & LeCun, 2025). Empowered by the strong self-supervision capability, JEPA brings significant advancements to various modalities such as video, audio, 3D vision, point cloud, etc. (Bardes et al., 2024; Fei et al., 2023; Hu et al., 2024; Saito et al., 2025), as well as multimodal learning (Lei et al., 2024; Chen et al., 2026; Cornelissen et al., 2026). Further, through large-scale pretraining, it is shown that JEPA is capable of fulfilling a world model (Ha & Schmidhuber, 2018b) to enable understanding, predicting, and planning in realistic environments (Assran et al., 2025; Mur-Labadia et al., 2026; Zhang et al., 2026a; Maes et al., 2026). Recently, the significance of JEPA has extended beyond AI, touching on realistic and fundamental applications such as medical imaging, spatial transcriptomics, antibody engineering, healthcare, and particle physics (Zhou et al., 2026a; Rabinowitz et al., 2026; Zhou et al., 2026b; Birk et al., 2026; Shihabi et al., 2026; Kim, 2026; Letellier et al., 2026). The proposed A-JEPA serves as a backbone with provable identifiability guarantees and might be readily adapted to a broad range of JEPA variants. Under the mild conditions specified in Assumptions 3.1 and 3.2, A-JEPA remains applicable to practical settings involving model scaling, dataset scaling, and multimodal learning. Overall, this work advances JEPA research by combining broadly generalizable empirical improvements with rigorous theoretical guarantees, while providing valuable insights into action-based JEPA training.

Recent theoretical analysis of JEPA provides latent-variable recovery guarantees up to an orthogonal transformation (Klindt et al., 2026), under the assumption of Gaussian latent dynamics. Related work on controlled world models considers Gaussian latent states and establishes joint state and transition identifiability under spectral separation and non-degenerate conditional action excitation, again up to an orthogonal transformation (Zhang et al., 2026b). Another line of work removes the Gaussian latent-distribution assumption by introducing a physics-grounded symbolic world model, under the assumption that a predefined symbolic atom library contains a causal basis for the underlying world generator (Dobrin & Chmiel, 2026). In contrast, our analysis assumes neither a Gaussian form for the latent transition mechanisms nor a known symbolic basis, and establishes component-wise identifiability for general nonparametric causal transitions.

Nonlinear ICA. This work is related to nonlinear ICA (Hyvarinen & Morioka, 2016; Hyvarinen et al., 2019; Hyvarinen & Morioka, 2017; Khemakhem et al., 2020), which studies the identifiability of latent components from nonlinear observations, typically under assumptions of mutual independence among latents. In particular, two key assumptions in our analysis, Assumptions 3.1 and 3.2, are closely related to the invertible observation mapping and variability conditions commonly used in nonlinear ICA identifiability theory. Our setting differs in that the latent variables are causal states whose transition mechanisms may depend on one another through a temporal causal graph, rather than mutually independent latent components, which introduces additional structure and also solution ambiguity.

Causal Representation Learning This work is also related to causal representation learning (CRL) (Schölkopf et al., 2021), which aims to recover latent causal variables from observations. A broad line of work has established identifiability results by exploiting temporal dependence (Yao et al., 2021; Song et al., 2023; Chen et al., 2024; Li et al., 2025; Lippe et al., 2022; Lachapelle et al., 2022). Different from most of these works, which typically establish identifiability through likelihood matching in observational space, we study identifiability within the JEPA framework, where representations are identified through action-conditioned prediction in latent space. Further, another line of work exploits interventions or multiple environments for identifiability (Brehmer et al., 2022; von Kügelgen et al., 2023; Ahuja et al., 2023; Seigal et al., 2022; Buchholz et al., 2023; Varici et al., 2023; Shen et al., 2022; Liu et al., 2024; 2026). Many of these works focus on instantaneous causal structures. while we consider temporal causal states and exploit action-induced variation in these mechanisms for identifiability.

Model-Based Reinforcement Learning Action-conditioned latent prediction has been extensively studied in model-based reinforcement learning. PlaNet and Dreamer learn latent dynamics by

matching an action-conditioned predictive prior to an observation-conditioned posterior, enabling planning or policy learning through imagined latent trajectories (Hafner et al., 2019b;a; 2025). Other world-model approaches also learn action-conditioned latent dynamics without requiring full observation reconstruction: MuZero learns latent dynamics specialized for predicting quantities relevant to planning (Schrittwieser et al., 2020), while TPC, EfficientZero, and TD-MPC/TD-MPC2 employ temporal prediction, latent consistency, or task-oriented latent dynamics for planning and control (Nguyen et al., 2021; Ye et al., 2021; Hansen et al., 2022; 2024). Closely related self-predictive approaches such as SPR explicitly predict future latent representations through learned transition models (Schwarzer et al., 2020), while reconstruction-free representation-learning methods based on bisimulation seek task-relevant latent abstractions without pixel decoding (Zhang et al., 2020). More recently, DreamerPro and R2-Dreamer replace observation reconstruction with self-supervised representation regularization (Deng et al., 2022; Morihira et al., 2026). These works demonstrate a broader pattern of learning action-conditioned latent dynamics together with mechanisms that preserve task-relevant or non-degenerate representations. However, they primarily target predictive and control performance rather than establishing recovery of the underlying causal state up to component-wise transformations. Our results address this identifiability question for this broader family of predictive representation objectives.

B LIMITATIONS

One of the main contributions of this work is establishing a component-wise identifiability result under the JEPA framework. As in existing identifiability analyses, our result relies on several assumptions on the underlying generative process. In particular, we assume that the observation mapping is smooth and invertible, such that information about the latent state is preserved through the observation process. Closely related assumptions are standard in nonlinear ICA and causal representation learning (Hyvarinen & Morioka, 2016; Hyvarinen et al., 2019; Khemakhem et al., 2020; von Kügelgen et al., 2023; Liu et al., 2026). We further require sufficient action-induced variation in the latent transition mechanisms, formalized through a full-rank variability condition. This type of variability condition also has a well-established precedent: it originates from identifiability analyses for nonlinear ICA (Hyvarinen & Morioka, 2016; Hyvarinen et al., 2019; Khemakhem et al., 2020) and has subsequently been adapted, in different forms, to causal representation learning (Zhang et al., 2024; Liu et al., 2026; Yao et al., 2021).

As in other identifiability analyses of latent-variable models, directly verifying these assumptions in real-world settings is inherently challenging, because the underlying latent generative process is unobserved. Nevertheless, our controlled experiments show strong recovery under the matched setting and robustness when observation invertibility, action-induced mechanism variation, and transition-noise specification are moderately perturbed. Developing weaker identifiability conditions and tighter connections between the general objective and practical estimators remains an important direction for future work.

C Proof of Theorem 2.1

Proof. For brevity, let $\mathbf{c}_t := (\mathbf{z}_{t-1}, \mathbf{a}_{t-1})$. For a given encoder $\mathbf{h}$, $Q_{\mathbf{h}}(\mathbf{z}_t \mid \mathbf{c}_t)$ denotes the conditional transition distribution induced in the representation space. Then, the conditional negative log-likelihood in the proposed objective Eq. (4) can be decomposed as

$$\begin{aligned}-\mathbb{E}\left[\log p_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{c}_t)\right] =& -\mathbb{E}_{\mathbf{c}_t}\mathbb{E}_{\mathbf{z}_t\sim Q_{\mathbf{h}}(\cdot\mid\mathbf{c}_t)}\left[\log Q_{\mathbf{h}}(\mathbf{z}_t \mid \mathbf{c}_t)\right] + \mathbb{E}_{\mathbf{c}_t}\mathbb{E}_{\mathbf{z}_t\sim Q_{\mathbf{h}}(\cdot\mid\mathbf{c}_t)}\left[\log \frac{Q_{\mathbf{h}}(\mathbf{z}_t \mid \mathbf{c}_t)}{p_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{c}_t)}\right] \\ =& H(\mathbf{z}_t \mid \mathbf{c}_t) + \mathbb{E}_{\mathbf{c}_t}\left[D_{\mathrm{KL}}\left(Q_{\mathbf{h}}(\cdot \mid \mathbf{c}_t) \| p_{\boldsymbol{\phi}}(\cdot \mid \mathbf{c}_t)\right)\right]. \end{aligned} \tag{16}$$

Substituting Eq. (16) into the proposed objective Eq. (4) and using $H(\mathbf{z}_t \mid \mathbf{c}_t) = H(\mathbf{z}_t) - I(\mathbf{z}_t; \mathbf{c}_t)$, we obtain

$$\mathcal{L}_\lambda(\mathbf{h}, \boldsymbol{\phi}) = -I(\mathbf{z}_t; \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) + \mathbb{E}_{\mathbf{c}_t}\left[D_{\mathrm{KL}}\left(Q_{\mathbf{h}}(\cdot \mid \mathbf{c}_t) \,\|\, p_{\boldsymbol{\phi}}(\cdot \mid \mathbf{c}_t)\right)\right] + (\lambda - 1)\big(-H(\mathbf{z}_t)\big). \tag{17}$$

Since $\mathbf{z}_t = \mathbf{h}(\mathbf{g}(\mathbf{s}_t))$ is a deterministic function of $\mathbf{s}_t$, and thus $(\mathbf{z}_{t-1}, \mathbf{a}_{t-1})$ is a deterministic function of $(\mathbf{s}_{t-1}, \mathbf{a}_{t-1})$, applying the data processing inequality to the two arguments gives

$$I(\mathbf{z}_t; \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) \le I(\mathbf{s}_t; \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) \le I(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}). \tag{18}$$

Moreover, because $\mathbf{h} : \mathcal{X} \to (0,1)^d$, the differential entropy of $\mathbf{z}_t$ satisfies

$$H(\mathbf{z}_t) \le \log \mathrm{Vol}((0,1)^d) = 0. \tag{19}$$

Together with the non-negativity of the KL divergence and the data-processing inequality in Eq. (18), Eq. (17) gives

$$\begin{aligned}\mathcal{L}_\lambda(\mathbf{h}, \boldsymbol{\phi}) + I(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) =& \underbrace{I(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) - I(\mathbf{z}_t; \mathbf{z}_{t-1}, \mathbf{a}_{t-1})}_{\ge 0} \\ &+ \underbrace{\mathbb{E}_{\mathbf{c}_t}\left[D_{\mathrm{KL}}\left(Q_{\mathbf{h}}(\cdot \mid \mathbf{c}_t) \| p_{\boldsymbol{\phi}}(\cdot \mid \mathbf{c}_t)\right)\right]}_{\ge 0} \\ &+ \underbrace{(\lambda - 1)\big(-H(\mathbf{z}_t)\big)}_{\ge 0}, \end{aligned} \tag{20}$$

where the last term is non-negative when $\lambda \ge 1$. As a result, we have:

$$\mathcal{L}_\lambda(\mathbf{h}, \boldsymbol{\phi}) \ge -I(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}). \tag{21}$$

When this lower bound is attainable, any global minimizer attaining the bound must make all three non-negative terms in Eq. (20) equal to zero. That is,

$$I(\mathbf{z}_t; \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = I(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}), \tag{22}$$

and

$$p_{\boldsymbol{\phi}^\star}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = Q_{\mathbf{h}^\star}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) \quad \text{a.e.} \tag{23}$$

This completes the proof. As an additional consequence, when $\lambda > 1$, Eq. (20) also implies $H(\mathbf{z}_t) = 0$. Since the maximum differential entropy over $(0,1)^d$ is 0, attained uniquely by the uniform distribution, we obtain $\mathbf{z}_t \sim \mathrm{Unif}((0,1)^d)$. □

D PROOF OF THEOREM 3.3

Proof sketch. The proof proceeds in three steps.

Step I: Local invertibility. Using the information-preservation result in Theorem 2.1 together with sufficient action variability, we show that the Jacobian of $\boldsymbol{r} = \boldsymbol{h} \circ \boldsymbol{g}$ has full rank everywhere, which implies that $\boldsymbol{r}$ is locally invertible.

Step II: Ruling out cross-component mixing. Using the local inverse of $\boldsymbol{r}$, together with the sufficient action variability, i.e., Assumption 3.2, we show that each row of the Jacobian of the local inverse contains at most one nonzero entry.

Step III: Component-wise identifiability. Combining Steps I and II, the Jacobian of the local inverse must be a generalized permutation matrix. Under regularity of the latent domain, the permutation pattern is fixed throughout the domain, yielding component-wise invertible transformations of the latent causal states.

Proof. **Step I: Local invertibility of $\boldsymbol{r}$.** Let $\boldsymbol{r} = \boldsymbol{h} \circ \boldsymbol{g}$, $\mathbf{z}_t = \boldsymbol{r}(\mathbf{s}_t)$. By Theorem 2.1, specifically Eq. (22), every global minimizer satisfies:

$$I(\mathbf{z}_t; \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = I(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}). \tag{24}$$

Since $\mathbf{z}_t$ is a deterministic function of $\mathbf{s}_t$, and accordingly, $(\mathbf{z}_{t-1}, \mathbf{a}_{t-1})$ is a deterministic function of $(\mathbf{s}_{t-1}, \mathbf{a}_{t-1})$, the data-processing inequality gives

$$I(\mathbf{z}_t; \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) \leq I(\mathbf{z}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) \leq I(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}). \tag{25}$$

Together with Eq. (24), both inequalities must be equalities, as follows:

$$I(\mathbf{z}_t; \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = I(\mathbf{z}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) = I(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}). \tag{26}$$

By the chain rule for mutual information, i.e., $I(X; Z \mid Y) = I(X, Y; Z) - I(Y; Z)$, together with $\mathbf{z}_t = \boldsymbol{r}(\mathbf{s}_t)$, we have

$$\begin{aligned} I(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1} \mid \mathbf{z}_t) &= I(\mathbf{s}_t, \mathbf{z}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) - I(\mathbf{z}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) && (27)\\ &= \underbrace{I(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) - I(\mathbf{z}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1})}_{=0 \text{ by Eq. (26)}} && (28)\\ &= 0. && (29) \end{aligned}$$

Here, $I(\mathbf{s}_t, \mathbf{z}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1})$ in Eq. (27) is equal to $I(\mathbf{s}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1})$ in Eq. (28), because $\mathbf{z}_t = \boldsymbol{r}(\mathbf{s}_t)$ is a deterministic function of $\mathbf{s}_t$. Similarly, since $\mathbf{z}_{t-1} = \boldsymbol{r}(\mathbf{s}_{t-1})$, we have

$$\begin{aligned} I(\mathbf{z}_t; \mathbf{s}_{t-1} \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) &= I(\mathbf{z}_t; \mathbf{s}_{t-1}, \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) - I(\mathbf{z}_t; \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) && (30)\\ &= \underbrace{I(\mathbf{z}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) - I(\mathbf{z}_t; \mathbf{z}_{t-1}, \mathbf{a}_{t-1})}_{=0 \text{ by Eq. (26)}} && (31)\\ &= 0. && (32) \end{aligned}$$

Similarly, $I(\mathbf{z}_t; \mathbf{s}_{t-1}, \mathbf{z}_{t-1}, \mathbf{a}_{t-1})$ in Eq. (30) is equal to $I(\mathbf{z}_t; \mathbf{s}_{t-1}, \mathbf{a}_{t-1})$ in Eq. (31), because $\mathbf{z}_{t-1} = \boldsymbol{r}(\mathbf{s}_{t-1})$ is a deterministic function of $\mathbf{s}_{t-1}$.

Eq. (29) implies

$$p(\mathbf{s}_{t-1}, \mathbf{a}_{t-1} \mid \mathbf{s}_t, \mathbf{z}_t) = p(\mathbf{s}_{t-1}, \mathbf{a}_{t-1} \mid \mathbf{z}_t). \tag{33}$$

Since $\mathbf{z}_t = \boldsymbol{r}(\mathbf{s}_t)$ is deterministic given $\mathbf{s}_t$, the left-hand side in Eq. (33) equals $p(\mathbf{s}_{t-1}, \mathbf{a}_{t-1} \mid \mathbf{s}_t)$, yielding

$$p(\mathbf{s}_{t-1}, \mathbf{a}_{t-1} \mid \mathbf{s}_t) = p(\mathbf{s}_{t-1}, \mathbf{a}_{t-1} \mid \mathbf{z}_t). \tag{34}$$

Applying Bayes' rule to both sides gives

$$\frac{p(\mathbf{s}_t \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) p(\mathbf{s}_{t-1}, \mathbf{a}_{t-1})}{p(\mathbf{s}_t)} = \frac{p(\mathbf{z}_t \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) p(\mathbf{s}_{t-1}, \mathbf{a}_{t-1})}{p(\mathbf{z}_t)}. \tag{35}$$

Canceling the common term yields

$$\frac{p(\mathbf{s}_t \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1})}{p(\mathbf{s}_t)} = \frac{p(\mathbf{z}_t \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1})}{p(\mathbf{z}_t)}. \tag{36}$$

Similarly, Eq. (32) implies $p(\mathbf{z}_t \mid \mathbf{s}_{t-1}, \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = p(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1})$. Since $\mathbf{z}_{t-1} = \boldsymbol{r}(\mathbf{s}_{t-1})$ is deterministic given $\mathbf{s}_{t-1}$, the left-hand side reduces to $p(\mathbf{z}_t \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1})$. Hence,

$$p(\mathbf{z}_t \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) = p(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = Q_{\boldsymbol{h}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}). \tag{37}$$

Substituting Eq. (37) into Eq. (36) gives

$$\frac{p(\mathbf{s}_t \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1})}{p(\mathbf{s}_t)} = \frac{Q_{\boldsymbol{h}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1})}{p(\mathbf{z}_t)}. \tag{38}$$

We next exploit the action variability in Assumption 3.2. Fix $(\mathbf{s}_t, \mathbf{s}_{t-1})$ and consider the corresponding $2d+1$ actions $\mathbf{a}_{t-1}^{(0)}, \ldots, \mathbf{a}_{t-1}^{(2d)}$. For each $\ell = 1, \ldots, 2d$, taking logarithms of Eq. (38) and subtracting the equation under $\mathbf{a}_{t-1}^{(0)}$ from that under $\mathbf{a}_{t-1}^{(\ell)}$ gives

$$\begin{aligned} &\log p(\mathbf{s}_t \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1}^{(\ell)}) - \log p(\mathbf{s}_t \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1}^{(0)}) \\ &\quad = \log Q_{\boldsymbol{h}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}^{(\ell)}) - \log Q_{\boldsymbol{h}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}^{(0)}). \end{aligned} \tag{39}$$

Using the latent transition factorization, $\log p(\mathbf{s}_t \mid \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) = \sum_{i=1}^{d} q_i$, where q_i is defined in Assumption 3.2. Differentiating Eq. (39) with respect to $\mathbf{s}_t$ and using $\mathbf{z}_t = \boldsymbol{r}(\mathbf{s}_t)$ gives

$$\begin{pmatrix} q_1'(\mathbf{a}_{t-1}^{(\ell)}) - q_1'(\mathbf{a}_{t-1}^{(0)}) \\ \vdots \\ q_d'(\mathbf{a}_{t-1}^{(\ell)}) - q_d'(\mathbf{a}_{t-1}^{(0)}) \end{pmatrix} = J_{\boldsymbol{r}}(\mathbf{s}_t)^\top \Big[\nabla_{\mathbf{z}_t} \log Q_{\boldsymbol{h}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}^{(\ell)}) - \nabla_{\mathbf{z}_t} \log Q_{\boldsymbol{h}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}^{(0)}) \Big], \tag{40}$$

where $q_i' = \frac{\partial q_i}{\partial s_{t,i}}$, and $\mathbf{z}_{t-1}$ is fixed when differentiating with respect to $\mathbf{s}_t$.

For brevity, with $(\mathbf{s}_t, \mathbf{s}_{t-1})$ fixed, write $\mathbf{w}(\mathbf{a}_{t-1}) = \mathbf{w}(\mathbf{s}_t, \mathbf{s}_{t-1}, \mathbf{a}_{t-1})$. By Assumption 3.2, the matrix

$$\boldsymbol{L} = \Big[\boldsymbol{w}(\mathbf{a}_{t-1}^{(1)}) - \boldsymbol{w}(\mathbf{a}_{t-1}^{(0)}), \ldots, \boldsymbol{w}(\mathbf{a}_{t-1}^{(2d)}) - \boldsymbol{w}(\mathbf{a}_{t-1}^{(0)}) \Big] \tag{41}$$

is invertible, where $\boldsymbol{w} = (q_1', \ldots, q_d', q_1'', \ldots, q_d'')^\top$.

Since $\boldsymbol{L}$ is invertible, the submatrix formed by its first d rows has rank d. Its columns are

$$\mathbf{v}_\ell = \begin{pmatrix} q_1'(\mathbf{a}_{t-1}^{(\ell)}) - q_1'(\mathbf{a}_{t-1}^{(0)}) \\ \vdots \\ q_d'(\mathbf{a}_{t-1}^{(\ell)}) - q_d'(\mathbf{a}_{t-1}^{(0)}) \end{pmatrix}, \qquad \ell = 1, \ldots, 2d, \tag{42}$$

and therefore span $\mathbb{R}^d$. By Eq. (40), each $\mathbf{v}_\ell$ can be written as $\mathbf{v}_\ell = J_{\boldsymbol{r}}(\mathbf{s}_t)^\top \mathbf{u}_\ell$ for some $\mathbf{u}_\ell \in \mathbb{R}^d$. Hence, stacking these vectors column-wise gives

$$\boldsymbol{V} = J_{\boldsymbol{r}}(\mathbf{s}_t)^\top \boldsymbol{U}, \tag{43}$$

with $\operatorname{rank}(\boldsymbol{V}) = d$. Therefore,

$$d = \operatorname{rank}(\boldsymbol{V}) \le \operatorname{rank}(J_{\boldsymbol{r}}(\mathbf{s}_t)^\top) \le d, \tag{44}$$

which implies

$$\operatorname{rank} J_{\boldsymbol{r}}(\mathbf{s}_t) = d. \tag{45}$$

Since the above argument applies throughout the latent support, $J_{\boldsymbol{r}}(\mathbf{s}_t)$ has full rank everywhere. By the inverse function theorem, $\boldsymbol{r}$ is locally invertible around every point in the latent support.

Step II: Ruling out cross-component mixing. Consider an arbitrary neighborhood on which a local inverse $\boldsymbol{T}$ of $\boldsymbol{r}$ is defined. Using $\mathbf{s}_t = \boldsymbol{T}(\mathbf{z}_t)$, Eq. (39) becomes

$$\sum_{i=1}^{d} \left[q_i(\mathbf{a}_{t-1}^{(\ell)}) - q_i(\mathbf{a}_{t-1}^{(0)}) \right] = \log Q_{\boldsymbol{h}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}^{(\ell)}) \quad - \log Q_{\boldsymbol{h}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}^{(0)}), \tag{46}$$

where the left-hand side is evaluated at $s_{t,i} = T_i(\mathbf{z}_t)$.

At the lower-bound-attaining global optimum, Theorem 2.1 gives $Q_{\boldsymbol{h}} = p_{\boldsymbol{\phi}}$ a.e. Moreover, the transition model factorizes across representation components:

$$p_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \prod_{j=1}^{d} p_{\boldsymbol{\phi},j}(z_{t,j} \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}). \tag{47}$$

Hence, for any $j \neq k$,

$$\frac{\partial^2}{\partial z_{t,j} \partial z_{t,k}} \log Q_{\boldsymbol{h}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = 0. \tag{48}$$

Taking the mixed derivative of Eq. (46) with respect to $z_{t,j}$ and $z_{t,k}$ gives

$$\sum_{i=1}^{d} \left[\left(q_i'(\mathbf{a}_{t-1}^{(\ell)}) - q_i'(\mathbf{a}_{t-1}^{(0)}) \right) \frac{\partial^2 T_i}{\partial z_{t,j} \partial z_{t,k}} + \left(q_i''(\mathbf{a}_{t-1}^{(\ell)}) - q_i''(\mathbf{a}_{t-1}^{(0)}) \right) \frac{\partial T_i}{\partial z_{t,j}} \frac{\partial T_i}{\partial z_{t,k}} \right] = 0. \tag{49}$$

For each $j \neq k$, define

$$\boldsymbol{v}_{jk} = \left(\frac{\partial^2 T_1}{\partial z_{t,j} \partial z_{t,k}}, \ldots, \frac{\partial^2 T_d}{\partial z_{t,j} \partial z_{t,k}}, \frac{\partial T_1}{\partial z_{t,j}} \frac{\partial T_1}{\partial z_{t,k}}, \ldots, \frac{\partial T_d}{\partial z_{t,j}} \frac{\partial T_d}{\partial z_{t,k}} \right)^{\top}. \tag{50}$$

Evaluating Eq. (49) for $\ell = 1, \ldots, 2d$ yields

$$\boldsymbol{L}^{\top} \boldsymbol{v}_{jk} = 0. \tag{51}$$

Since $\boldsymbol{L}$ is invertible, $\boldsymbol{v}_{jk} = 0$. In particular,

$$\frac{\partial T_i}{\partial z_{t,j}} \frac{\partial T_i}{\partial z_{t,k}} = 0, \qquad \forall i, \quad j \neq k. \tag{52}$$

Thus, for any fixed i, two distinct partial derivatives of T_i cannot be nonzero simultaneously. Hence, each row of $J_{\boldsymbol{T}}(\mathbf{z}_t)$ contains at most one nonzero entry.

Step III: Component-wise identifiability. By Step I, $\boldsymbol{r}$ is locally invertible, so on every neighborhood where $\boldsymbol{T} = \boldsymbol{r}^{-1}$ is defined, $J_{\boldsymbol{T}}(\mathbf{z}_t)$ is nonsingular. Step II shows that each row of $J_{\boldsymbol{T}}(\mathbf{z}_t)$ contains at most one nonzero entry. Nonsingularity rules out zero rows, so each row contains exactly one nonzero entry. Since the matrix is square and nonsingular, each column also contains exactly one nonzero entry. Thus, $J_{\boldsymbol{T}}(\mathbf{z}_t)$ is a generalized permutation matrix. Its inverse $J_{\boldsymbol{r}}(\mathbf{s}_t) = J_{\boldsymbol{T}}(\mathbf{z}_t)^{-1}$ is also a generalized permutation matrix. Hence, locally, each component r_i depends on exactly one latent coordinate.

Since $J_{\boldsymbol{r}}$ is continuous and nonsingular, its generalized-permutation pattern is locally constant. On a connected latent domain, this pattern is constant throughout the domain. Hence, there exists a fixed permutation π such that, for all $i = 1, \ldots, d$,

$$\frac{\partial r_i}{\partial s_j} = 0 \quad \text{for } j \neq \pi(i), \qquad \frac{\partial r_i}{\partial s_{\pi(i)}} \neq 0. \tag{53}$$

On a connected product domain, Eq. (53) implies that each r_i depends only on $s_{\pi(i)}$. Thus, there exist one-dimensional functions $\rho_1, \ldots, \rho_d$ such that

$$z_{t,i} = \rho_i(s_{t,\pi(i)}), \qquad i = 1, \ldots, d. \tag{54}$$

Furthermore, $\rho_i'(s_{t,\pi(i)}) = \frac{\partial r_i}{\partial s_{\pi(i)}} \neq 0$. Since each one-dimensional latent domain is connected and ρ_i' is continuous and nowhere zero, ρ_i is strictly monotone and invertible. Consequently,

$$\mathbf{z}_t = \left(\rho_1(s_{t,\pi(1)}), \ldots, \rho_d(s_{t,\pi(d)}) \right), \tag{55}$$

which establishes component-wise identifiability up to permutation and component-wise invertible transformations. □

E Negative-Sample Contrastive Approximation to Entropy

We provide a theoretical interpretation of the Negative-Sample contrastive term used in Eq. (15). The argument follows the kernel-density interpretation of contrastive uniformity in Wang & Isola (2020), while adapting it to our action-conditioned transition likelihood.

Consider an isotropic Gaussian transition model

$$p_{\phi}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \mathcal{N}\left(\mathbf{z}_t; \boldsymbol{\mu}_{\phi}(\mathbf{z}_{t-1}, \mathbf{a}_{t-1}), \sigma^2 \mathbf{I}\right). \tag{56}$$

Its negative log-likelihood is

$$\ell_{\phi}(\mathbf{z} \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \frac{1}{2\sigma^2} \left\|\mathbf{z} - \boldsymbol{\mu}_{\phi}(\mathbf{z}_{t-1}, \mathbf{a}_{t-1})\right\|_2^2 + \frac{d}{2} \log(2\pi\sigma^2). \tag{57}$$

Proposition E.1 (Entropy interpretation of the contrastive term)**.** *Let $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(K)}$ be i.i.d. samples from the marginal representation distribution $p_{\mathbf{z}}$. Suppose that the transition prediction is locally accurate, such that $\boldsymbol{\mu}_{\phi}(\mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \mathbf{z}_t$. Define*

$$\mathcal{U}_K = \log\left[\frac{1}{K} \sum_{k=1}^{K} \exp\left(-\frac{\ell_{\phi}(\mathbf{z}^{(k)} \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1})}{\tau}\right)\right]. \tag{58}$$

Then, as $K \to \infty$,

$$\mathcal{U}_K \longrightarrow \log\left[(p_{\mathbf{z}} * \varphi_h)(\mathbf{z}_t)\right] + C_{\sigma,\tau}, \qquad h^2 = \tau\sigma^2, \tag{59}$$

where φ_h is the Gaussian kernel with bandwidth h and $C_{\sigma,\tau}$ is independent of the representation distribution. Consequently,

$$\mathbb{E}_{\mathbf{z}_t}[\mathcal{U}_K] \longrightarrow -H_h(\mathbf{z}_t) + C_{\sigma,\tau}, \tag{60}$$

where

$$H_h(\mathbf{z}_t) := -\mathbb{E}_{\mathbf{z}_t} \log\left[(p_{\mathbf{z}} * \varphi_h)(\mathbf{z}_t)\right] \tag{61}$$

is the Gaussian-kernel resubstitution entropy. Under the standard regularity conditions for kernel-density estimation, as $h \to 0$,

$$H_h(\mathbf{z}_t) \longrightarrow H(\mathbf{z}_t). \tag{62}$$

Thus, up to a constant, minimizing the negative-sample term asymptotically corresponds to maximizing the representation entropy.

Proof. Under $\boldsymbol{\mu}_{\phi}(\mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \mathbf{z}_t$,

$$\begin{aligned} \exp\left(-\frac{\ell_{\phi}(\mathbf{z}^{(k)} \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1})}{\tau}\right) &= C'_{\sigma,\tau} \exp\left(-\frac{\|\mathbf{z}^{(k)} - \mathbf{z}_t\|_2^2}{2\tau\sigma^2}\right) \\ &= C_{\sigma,\tau} \varphi_h(\mathbf{z}^{(k)} - \mathbf{z}_t), \qquad h^2 = \tau\sigma^2, \end{aligned} \tag{63}$$

where $C_{\sigma,\tau}$ and $C'_{\sigma,\tau}$ do not depend on $\mathbf{z}_t$.

Therefore,

$$\mathcal{U}_K = \log\left[\frac{1}{K} \sum_{k=1}^{K} \varphi_h(\mathbf{z}^{(k)} - \mathbf{z}_t)\right] + C_{\sigma,\tau}. \tag{64}$$

The quantity inside the logarithm is precisely the Gaussian kernel-density estimator of $p_{\mathbf{z}}$ evaluated at $\mathbf{z}_t$. By the law of large numbers,

$$\frac{1}{K} \sum_{k=1}^{K} \varphi_h(\mathbf{z}^{(k)} - \mathbf{z}_t) \longrightarrow \int \varphi_h(\mathbf{z} - \mathbf{z}_t) p_{\mathbf{z}}(\mathbf{z})\, d\mathbf{z} = (p_{\mathbf{z}} * \varphi_h)(\mathbf{z}_t). \tag{65}$$

This establishes Eq. (59). Taking expectation over $\mathbf{z}_t$ gives

$$\mathbb{E}[\mathcal{U}_K] \longrightarrow -H_h(\mathbf{z}_t) + C_{\sigma,\tau}.$$

Finally, consistency of Gaussian kernel-density estimation gives $H_h(\mathbf{z}_t) \to H(\mathbf{z}_t)$ as the bandwidth vanishes under the usual regularity conditions. □

F IMPLEMENTATION OF A-JEPA

A key practical question in A-JEPA is how to learn the temporal DAG structure in the representation space. We first introduce a predefined causal order, which allows us to incorporate DAG structure through a fully connected candidate graph consistent with this order. Ideally, with sufficient data, a well-specified model, and global optimization, the learned transition mechanisms recover the true conditional dependencies, making non-parent candidate inputs functionally redundant. These ideal conditions are difficult to satisfy in practice, however, and redundant edges may not be clearly suppressed under finite-sample optimization. We therefore also introduce a sparse implementation with learnable edge gates and an ℓ_1 regularizer for explicit parent selection.

Predefined causal order in representation space. The component-wise identifiability result determines the latent causal variables only up to permutation. Consequently, the indices of the learned representation coordinates carry no predefined semantic meaning. For a latent transition graph that is acyclic over the latent components, there exists a topological ordering under which every non-self causal edge points from a lower-order component to a higher-order component. Since permutations of the learned coordinates represent equivalent identifiable solutions, we fix the coordinate order $z_1 \prec \cdots \prec z_d$ as a canonical representative, and denote the ordered coordinates by $z_1, \ldots, z_d$. This choice does not assume the ground-truth semantic labels of the latent variables; it only fixes one representative within the permutation-equivalent latent space, following (Liu et al., 2026).

For example, consider three latent variables with non-self causal relations $s_3 \to s_2$, $s_3 \to s_1$, and $s_2 \to s_1$. One valid topological ordering is (s_3, s_2, s_1). Suppose the learned representation instead recovers (s_1, s_3, s_2) up to component-wise invertible transformations. Since the identifiable solution is invariant to permutation, we may simply reorder the learned coordinates as (s_3, s_2, s_1). Under this canonical ordering, all non-self causal edges point from lower-order to higher-order coordinates. Thus, imposing such an ordering fixes only the otherwise arbitrary permutation of the learned representation, rather than introducing additional semantic information about the latent variables.

Implementation of A-JEPA. Based on this predefined order, we construct a fully connected candidate DAG in the representation space. For each child component $z_{t,i}$, all lower-order components $z_{t-1,j}$ with $j < i$ are treated as candidate non-self temporal parents, together with the self-transition $z_{t-1,i} \to z_{t,i}$. Accordingly, the candidate parent set is

$$\mathrm{pa}_i = \{1, \ldots, i-1\} \cup \{i\}. \tag{66}$$

Because all non-self edges follow the predefined order, the resulting graph is acyclic by construction.

Using this full candidate parent set, the Gaussian transition model in Eq. (13) becomes

$$p_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \prod_{i=1}^{d} \mathcal{N}\left(z_{t,i}; \mu_{\boldsymbol{\phi},i}(\mathbf{z}_{t-1,j\le i}, \mathbf{a}_{t-1}), \sigma^2_{\boldsymbol{\phi},i}(\mathbf{a}_{t-1})\right). \tag{67}$$

```
for each minibatch (x_{t-1}, a_{t-1}, x_t):
    z_{t-1} = h(x_{t-1})
    z_t     = h(x_t)
    for i = 1, ..., d:
        # candidate parents follow the predefined order
        parent_i = z_{t-1, 1:i}
        mu_i     = mean_phi_i(parent_i, a_{t-1})
        sigma_i  = std_phi_i(a_{t-1})
    # positive transition likelihood
    nll_pos = GaussianNLL(z_t,{mu_i, sigma_i}_{i=1}^d)
    # sample alternative future representations
    {z_t^(k)}_{k=1}^K = SampleNegatives(z_t)
    for k = 1, ..., K:
        nll_neg[k] = GaussianNLL(z_t^(k),{mu_i, sigma_i}_{i=1}^d)
    # A-JEPA objective
    L = alpha*nll_pos + (1-alpha)*log((1/K)*sum_k exp(-nll_neg[k]/tau))
    update h, phi by minimizing L
```

Listing 1: A-JEPA with a fully connected candidate DAG.

Implementation of A-JEPA with Sparse Parent Selection. In practice, finite samples and imperfect optimization make it difficult to determine whether a fitted transition mechanism is functionally independent of a redundant candidate parent. We therefore make the parent selection explicit by introducing a learnable scalar gate g_{ji} for each candidate non-self edge $z_{t-1,j} \to z_{t,i}$ with $j < i$. The self-transition $z_{t-1,i} \to z_{t,i}$ is always retained.

For the i-th transition mechanism, we define the gated parent input as

$$\widetilde{\mathbf{z}}^{(i)}_{t-1,1:i} = \left(g_{1i} z_{t-1,1}, \ldots, g_{(i-1)i} z_{t-1,i-1}, z_{t-1,i}\right). \tag{68}$$

The Gaussian transition model is then implemented as

$$p_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \prod_{i=1}^{d} \mathcal{N}\left(z_{t,i}; \mu_{\boldsymbol{\phi},i}(\widetilde{\mathbf{z}}^{(i)}_{t-1,1:i}, \mathbf{a}_{t-1}), \sigma^2_{\boldsymbol{\phi},i}(\mathbf{a}_{t-1})\right). \tag{69}$$

To encourage sparse parent selection, we augment the practical A-JEPA objective in Eq. (15) with an ℓ_1 penalty on the non-self gates:

$$\mathcal{L}_{\text{sparse}} = \mathcal{L} + \lambda_{\text{sp}} \sum_{i=1}^{d} \sum_{j<i} |g_{ji}|, \tag{70}$$

where $\lambda_{\text{sp}} > 0$ controls the sparsity strength. After optimization, the estimated non-self parent set of $z_{t,i}$ is obtained by thresholding the learned gates:

$$\widehat{\text{pa}}^{z}_i = \{j < i : |g_{ji}| > \delta\} \cup \{i\}, \tag{71}$$

where δ is a fixed threshold.

```
for each minibatch (x_{t-1}, a_{t-1}, x_t):
    z_{t-1} <- h(x_{t-1})
    z_t     <- h(x_t)
    for i = 1, ..., d:
        P_i <- (g_{1i} z_{t-1,1},...,g_{(i-1)i} z_{t-1,i-1},z_{t-1,i})
        mu_i    <- mean_phi_i(P_i, a_{t-1})
        sigma_i <- std_phi_i(a_{t-1})
    L_pos <- GaussianNLL(z_t | z_{t-1}, a_{t-1})
    {z_t^(k)}_{k=1}^K <- sample alternative futures
    for k = 1, ..., K:
        L_neg^(k) <- GaussianNLL(z_t^(k) | z_{t-1},a_{t-1})
    L <- alpha*L_pos + (1-alpha) * log[(1/K) * sum_k exp(-L_neg^(k) / tau
    )]
         + lambda_sp * sum_{i} sum_{j<i} |g_{ji}|
    update h, phi, and {g_{ji}} by minimizing L
for each i:
    pa_hat_i <- {j < i : |g_{ji}| > delta} union {i}
```

Listing 2: A-JEPA with sparse parent selection.

G Simulation Details

This section provides additional details of the synthetic experiments. The data are generated according to the action-modulated Gaussian ANM in Sec. 4. The simulation is designed to control the main factors appearing in our theory, including the observation mapping and action-induced variation in the latent transition mechanisms.

Latent transition model. We generate a latent causal state $\mathbf{s}_t \in \mathbb{R}^d$ following a temporal DAG. For each random seed, the DAG is sampled once and kept fixed across all episodes. Each latent variable has a self-transition and may have additional parents from the previous state, following the known causal order. Each variable has at most three temporal parents including itself. The transition follows

$$s_{t,i} = f_i(\mathbf{s}_{t-1,\text{pa}_i}, \mathbf{a}^{\mu}_{t-1}) + \sigma_i(\mathbf{a}^{\sigma}_{t-1})\epsilon_{t,i}, \qquad \epsilon_{t,i} \sim \mathcal{N}(0,1), \tag{72}$$

where

$$\mathbf{a}_{t-1} = [\mathbf{a}^{\mu}_{t-1}, \mathbf{a}^{\sigma}_{t-1}], \qquad \mathbf{a}^{\mu}_{t-1}, \mathbf{a}^{\sigma}_{t-1} \in \mathbb{R}^d. \tag{73}$$

Thus, the default action dimension is $2d$. The first action block modulates the conditional mean, while the second modulates the conditional variance.

For a temporal causal edge $j \to i$, its action-dependent coefficient is

$$A_{ji}(\mathbf{a}^{\mu}_{t-1}) = \kappa G_{ji} \tanh\left(\mathbf{w}_{ji}^{\top} \mathbf{a}^{\mu}_{t-1}\right), \qquad \kappa = 0.5. \tag{74}$$

Self-transitions additionally have a fixed coefficient $\rho = 0.6$. The resulting mean is passed through a LeakyReLU nonlinearity. Action coordinates not associated with an active parent mechanism are masked out, so that the conditional mean is modulated only through the sampled temporal causal edges. The conditional standard deviation is

$$\sigma_i(\mathbf{a}^{\sigma}_{t-1}) = \sigma_{\min} + s_{\sigma} \operatorname{softplus}(a^{\sigma}_{t-1,i}), \tag{75}$$

where $\sigma_{\min} = 0.08$ and $s_{\sigma} = 1.0$.

Action settings. The action is fixed within each episode and varied across episodes. Therefore, different episodes correspond to different action-conditioned transition mechanisms. Unless otherwise specified, each episode contains $2{,}000$ samples and we generate $10d$ episodes. To study the effect of action variability, we vary the number of action settings. For $d = 5$, we use 5, 10, 20, 30, 40, and 50 different action settings.

Observation model. The latent state is mapped to the observation as

$$\mathbf{x}_t = \mathbf{g}(\mathbf{s}_t), \tag{76}$$

where $\mathbf{g}$ is constructed using random orthogonal transformations and LeakyReLU nonlinearities. The default mapping is invertible. To examine the role of this assumption, we progressively remove observed coordinates while keeping the latent dimension fixed, making the observation mapping increasingly non-invertible.

Network architecture. The used encoder is a three-layer MLP with hidden width 256, where each hidden block consists of Linear-BatchNorm-SiLU layers, followed by a linear projection to the d-dimensional representation space. The encoder output is mapped coordinate-wise to $(\delta, 1-\delta)$ using a sigmoid transform with $\delta = 10^{-4}$. The used transition model predicts a factorized Gaussian distribution over the learned coordinates. For each coordinate i, the conditional mean is produced by a separate two-layer MLP with hidden width 256, taking as input the previous representations of its candidate temporal parents together with the mean-action variables. The conditional variance is predicted by a separate coordinate-wise MLP from the action variables, with log-variance clamped to $[-1, 1]$.

Model and training. The encoder maps $\mathbf{x}_t$ to a learned representation $\mathbf{z}_t$. The learned transition model predicts a diagonal Gaussian distribution

$$p_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1}) = \prod_{i=1}^{d} \mathcal{N}\left(z_{t,i}; \mu_{\boldsymbol{\phi},i}, \sigma^2_{\boldsymbol{\phi},i}\right), \tag{77}$$

where the conditional mean and variance depend on the previous representation and action. In the main representation-recovery experiments, A-JEPA uses the maximally connected candidate DAG consistent with the predefined causal order, as described in Appendix F. For graph recovery, we use its sparse-gated variant and learn the active non-self parents from data.

We optimize the Gaussian contrastive objective in Eq. (15). Unless otherwise specified, we use $\alpha = 0.51$, $\tau = 1.0$, Adam with learning rate 10^{-4}, weight decay 10^{-5}, batch size 512, and 200 training epochs. We use 80% of the generated transitions for training and 20% for testing.

Evaluation. Latent state recovery is measured using component-wise Spearman MCC. We compute the absolute Spearman correlations between all learned and ground-truth latent coordinates, align them using the Hungarian algorithm, and average the matched correlations. This metric is consistent with the component-wise equivalence class in Theorem 3.3.

For graph recovery, we first align the learned coordinates using the same matching. We then threshold the learned parent gates and report precision, recall, F1, and structural Hamming distance (SHD) against the ground-truth temporal DAG. Self-transitions are excluded from graph evaluation because they are always included in the simulation.

Assumption and model-specification analysis. We consider three controlled changes to the default setting. First, we reduce action-induced variation by allowing actions to modulate only the conditional mean or only the conditional variance. Second, we replace Gaussian noise with Laplace, Student-t, or Uniform noise while keeping the same action-dependent variance. Third, we progressively remove observation coordinates to weaken the invertibility of the observation mapping. These experiments examine how latent-state recovery changes when the main assumptions or modeling choices are relaxed.

Computational resources. All experiments were conducted on NVIDIA A100 GPUs with 40 GB of memory, with different random seeds and experimental settings executed independently.

H Visual Benchmark Details

This section provides additional details of the four visual control benchmarks used in Sec. 5. These environments provide image observations and ground-truth physical state variables. We therefore use them to evaluate whether the learned representations align with interpretable state factors, preserve information about the underlying state, and support action-conditioned latent prediction.

Encoder and transition architecture. For all visual benchmarks, A-JEPA and the baselines use the same compact CNN encoder. Given an RGB observation resized to 224×224, the encoder applies four convolutional blocks with channel widths 32, 64, 128, and 256. Each block consists of a 5×5 convolution with stride 2 and padding 2, followed by BatchNorm and LeakyReLU. The resulting feature map is globally average pooled and projected to the d_z-dimensional representation space through a two-layer MLP with hidden width 512. Following the data convention used by the predictive baselines, we use a frame skip of 5. The five primitive actions between two encoded frames are concatenated and mapped to a d_z-dimensional action representation through a small action encoder.

A-JEPA uses a coordinate-wise transition model following a predefined latent order, which is without loss of generality due to the permutation indeterminacy in latent space, see the discussion in Sec. F. For each coordinate i, the conditional mean is predicted from the previous latent coordinates up to i together with the encoded action,

$$\mu_{t+1,i} = f_i(\mathbf{z}_{t,\leq i}, \mathbf{a}_t^z)\,, \tag{78}$$

where each f_i is a separate two-layer MLP with hidden width 256. The transition variance is predicted by a separate action-conditioned head, with one log-variance per latent coordinate and values clamped to $[-1, 1]$. Together, these heads define the factorized Gaussian transition model used by the A-JEPA objective.

Baselines and training. All methods take image observations as input and use the same CNN encoder and comparable training setup. A-JEPA is trained with the objective in Eq. (15) and the action-conditioned transition model described above. LeWM and PLDM use the same image encoder, action representation, latent dimension, frame skip, optimizer settings, batch size, and number of training epochs, while retaining their original predictive objectives and non-DAG transition models.

Unless otherwise specified, models are trained from scratch for 10 epochs using Adam with learning rate 10^{-4}, weight decay 10^{-5}, batch size 128, and history size 1. We reserve 10% of the official training split for validation and evaluate on the held-out test split. For A-JEPA, we use $\alpha = 0.51$ and $\tau = 1.0$. All reported results are averaged over three random seeds.

Evaluation metrics. We evaluate the learned representations from three complementary perspectives. Component-wise MCC measures alignment between individual learned coordinates and ground-truth physical state factors. We compute absolute Spearman correlations, perform Hungarian

matching, and average the matched scores. This metric is consistent with the component-wise equivalence class in Theorem 3.3.

Nonlinear state R^2 measures how much information about the underlying state is retained in the full representation, regardless of coordinate-wise alignment. Latent rollout R^2 evaluates action-conditioned prediction by recursively applying the learned transition model and comparing the predicted future representation with the encoder representation of the corresponding future observation.

TwoRoom. TWOROOM is a two-dimensional navigation environment where the action-driven state is the agent position,

$$\mathbf{s}_t = [x_t, y_t]. \tag{79}$$

The dataset observation also contains target coordinates, but we exclude them from MCC because they specify the episode goal/context rather than the evolving state. Concretely, we evaluate component recovery only on the two dynamic coordinates (x_t, y_t). For evaluation, we use $5{,}000$ held-out test

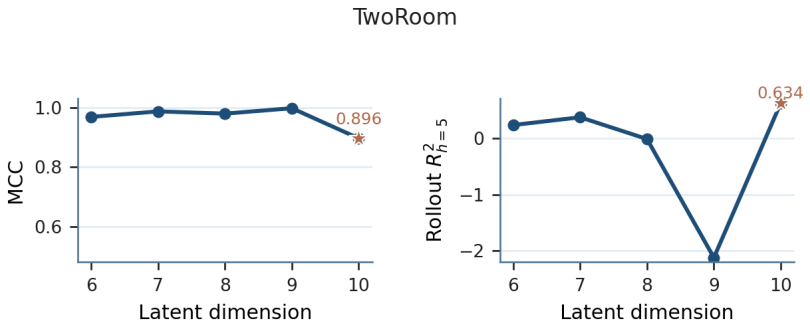


Figure 6: Latent-dimension selection on TWOROOM. We sweep $d_z \in \{6, 7, 8, 9, 10\}$ and evaluate both component-wise recovery (MCC) and rollout R^2 at horizon 5. We use $d_z = 10$ for the main experiments because it provides a strong balance between component recovery and multi-step transition prediction.

samples for each seed and compute the absolute Spearman correlation between every learned latent coordinate and each ground-truth state factor. We then apply Hungarian matching before averaging the matched scores. This accounts for both the permutation ambiguity and the component-wise monotone transformations allowed by our identifiability result. We sweep the latent dimension over $d_z \in \{6, 7, 8, 9, 10\}$. As shown in Figure 6, although smaller dimensions achieve high MCC, the $d_z = 10$ model gives the best rollout behavior among the sweep settings, with rollout $R^2_{h=5} = 0.6343$ in the dimension-selection run. We therefore use $d_z = 10$ for the main three-seed comparison in Figure 4, selecting the dimension by the joint criterion of component recovery and action conditioned dynamics quality.

Reacher. REACHER is a two-link articulated arm with two rotational state variables. Because joint angles are periodic, direct correlation with the raw angles can be misleading near the $-\pi/\pi$ boundary. Thus, we represent each angle q_j by its circular coordinates $(\sin q_j, \cos q_j)$. We measure component recovery over the two physical angular factors, namely the proximal joint angle q_0 and the distal joint angle q_1, treating each $(\sin q_j, \cos q_j)$ pair as one circular state factor rather than two independent scalar factors. For each learned coordinate z_i, we compute its best one-dimensional projection onto $(\sin q_j, \cos q_j)$ and use the resulting correlation for Hungarian matching across the two joint-angle factors. Evaluation is performed on $5{,}000$ held-out test samples. We sweep the latent dimension over $d_z \in \{6, 8, 10, 12, 14, 16\}$ and select the representation size by jointly considering component-wise recovery and multi-step prediction. As shown in Fig. 7, $d_z = 6$ achieves strong performance on both criteria, with MCC 0.926 and rollout $R^2_{h=5} = 0.973$, while increasing the latent dimension does not improve component recovery. We therefore use $d_z = 6$ for the main three-seed comparison.

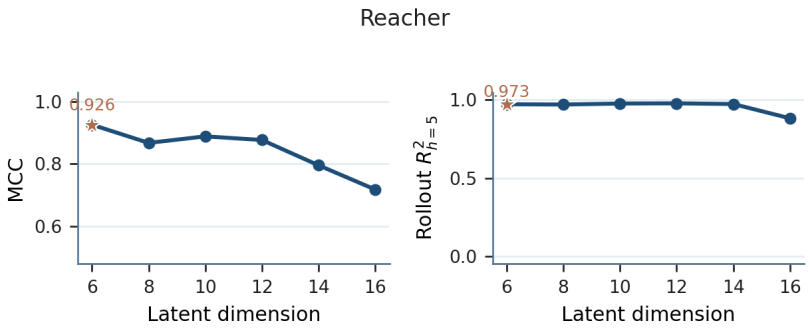


Figure 7: Latent-dimension selection on REACHER. We sweep $d_z \in \{6, 8, 10, 12, 14, 16\}$ and evaluate grouped circular MCC and rollout R^2 at horizon 5. We use $d_z = 6$ for the main experiments because it achieves strong component-wise recovery and multi-step prediction with the most compact representation.

PushT. PUSHT is a planar manipulation environment in which a controlled pusher moves a T-shaped block through contact. We evaluate component recovery using five physical state factors,

$$\mathbf{s}_t = [x_t^p, y_t^p, x_t^b, y_t^b, \theta_t^b], \tag{80}$$

where p and b denote the pusher and block, respectively. Since pusher motion is directly driven by the action, we additionally include two finite-difference motion factors,

$$\Delta x_t^p = x_{t+1}^p - x_t^p, \qquad \Delta y_t^p = y_{t+1}^p - y_t^p, \tag{81}$$

computed from adjacent pusher positions. The main MCC is evaluated over seven factors:

$$[x_t^p, y_t^p, x_t^b, y_t^b, \theta_t^b, \Delta x_t^p, \Delta y_t^p]. \tag{82}$$

Again, we select the representation dimension by jointly considering component-wise recovery and multi-step prediction. As shown in Fig. 8, the dimension sweep includes $d_z \in \{8, 9, 10, 11, 12, 16, 24\}$ for MCC, with rollout evaluation concentrated around the smaller dimensions. The $d_z = 10$ setting provides the strongest joint performance in the selection run, achieving MCC 0.620 and rollout $R^2_{h=5} = 0.919$. We use $d_z = 10$ for the main three-seed comparison reported in Figure 4.

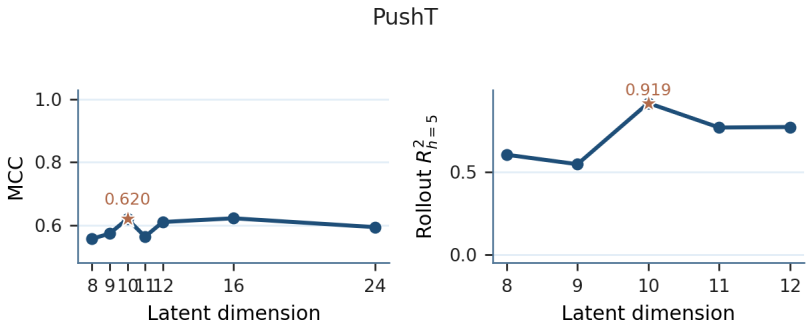


Figure 8: Latent-dimension selection on PUSHT. We evaluate both component-wise recovery and rollout R^2 at horizon 5. We use $d_z = 10$ for the main experiments because it provides the strongest joint component-recovery and multi-step prediction performance in the dimension-selection run.

OGBench Cube. OGBENCH CUBE is a robot manipulation environment involving end-effector motion, gripper actuation, and object manipulation. We evaluate component recovery on seven task-relevant physical factors,

$$\mathbf{s}_t = [x_t^{ee}, y_t^{ee}, z_t^{ee}, x_t^{obj}, y_t^{obj}, z_t^{obj}, g_t], \tag{83}$$

where ee denotes the end-effector, obj denotes the manipulated object, and g_t denotes the gripper opening. We use these factors rather than the full raw state vector because they directly describe the visible and action-relevant state of the manipulation task. For each seed, evaluation is performed on 5,000 held-out test samples. We sweep the latent dimension over $d_z \in 10, 12, 14, 16, 18$ and select

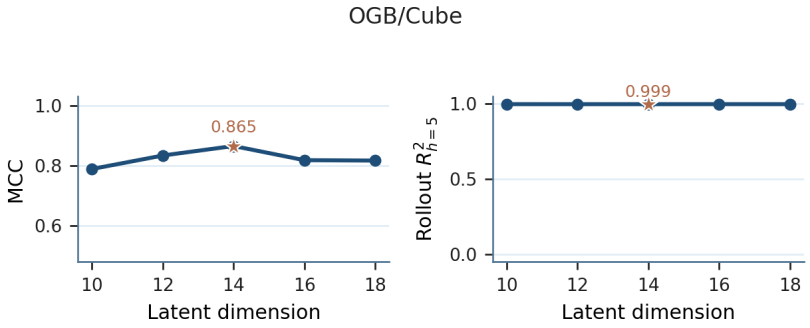


Figure 9: Latent-dimension selection on OGBENCH CUBE. We sweep $d_z \in \{10, 12, 14, 16, 18\}$ and evaluate component-wise recovery and rollout R^2 at horizon 5. We use $d_z = 14$ for the main experiments because it achieves the strongest joint recovery and multi-step prediction performance.

the representation size by jointly considering component-wise recovery and multi-step prediction. As shown in Fig. 9, $d_z = 14$ achieves the strongest overall performance, with MCC 0.865 and rollout $R^2_{h=5} = 0.999$, and is therefore used for the main three-seed comparison. At this dimension, A-JEPA strongly recovers the end-effector coordinates and horizontal object position, while object height and gripper opening remain more challenging. The learned transition is also highly stable, with rollout $R^2 = 0.9987$ at horizon 5 and 0.9940 ± 0.0051 at horizon 10.

I VISUAL INTERVENTION EXPERIMENTS ON REACHER

We further examine whether the component-wise semantics and directed latent dependencies learned by A-JEPA correspond to meaningful physical changes in the visual environment. We use REACHER for this analysis because its two-link articulated structure provides a simple setting in which the direction of physical dependencies can be interpreted clearly.

Our evaluation follows three steps. First, we assign physical semantics to individual learned coordinates using the grouped circular MCC described in Sec. H. Second, we inspect whether the directed edges learned between these coordinates are consistent with the physical structure of the system. Third, we intervene on the corresponding latent coordinates and propagate the intervention through the learned transition model to examine whether the resulting visual changes follow the learned direction of dependence.

Semantic alignment. The physical state of REACHER contains two angular factors: the proximal joint angle and the distal joint angle. Using grouped circular MCC, the learned coordinate z_0 is matched to the proximal joint factor, while z_3 is matched to the distal joint factor. Their corresponding matching scores are approximately 0.80 and 0.89 shown in Figure 10, respectively. This semantic assignment is performed before examining the learned graph or intervention results.

Learned directed dependency. After assigning semantics to the latent coordinates, we inspect the learned gated transition graph. As shown in Fig. 10, the model identifies a directed edge $z_0 \to z_3$, corresponding to a proximal-to-distal dependency. This relation is consistent with the physical structure of a two-link arm: changing the proximal joint changes the configuration of the upstream link and consequently affects the position and orientation of the downstream link, whereas the distal joint does not analogously determine the proximal joint.

Latent interventions. To further examine the direction of this learned dependency, we intervene on one latent coordinate at time t while keeping the remaining coordinates fixed. For a selected

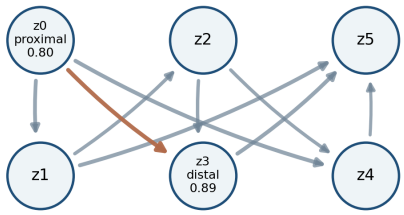


Figure 10: Learned latent transition graph on REACHER. MCC assigns z_0 to the proximal joint factor and z_3 to the distal joint factor. The learned graph contains the directed edge $z_0 \to z_3$, consistent with the proximal-to-distal dependency of the two-link arm.

coordinate z_i, we consider interventions of the form $z_{t,i}^{\mathrm{do}} = z_{t,i} + \delta\,\mathrm{Std}(z_i)$, where δ controls the intervention magnitude. The intervened latent state is then propagated through the learned action-conditioned transition model, $\hat{\mathbf{z}}_{t+1}^{\mathrm{do}} = T_{\boldsymbol{\phi}}(\mathbf{z}_t^{\mathrm{do}}, \mathbf{a}_t)$, and the predicted next representation is decoded back to image space. Thus, the visualization reflects the effect of the intervention after propagation through the learned transition dynamics, rather than simply decoding a manually perturbed latent coordinate.

Directional intervention results. The intervention responses are consistent with the learned edge $z_0 \to z_3$. As shown in Figure 11, when intervening on the proximal-aligned coordinate z_0, changing its value produces a clear change in the proximal joint and also induces a corresponding change in the distal arm segment. In contrast, intervening on the distal-aligned coordinate z_3 primarily changes the distal joint while leaving the proximal configuration largely unchanged. This asymmetric response is consistent with the learned proximal-to-distal direction: intervention on the parent affects both the parent and its downstream child, whereas intervention on the child produces a more localized change.

Taken together, these experiments provide three complementary checks on the learned latent structure: MCC assigns interpretable physical semantics to individual coordinates, the learned graph recovers a physically plausible directed dependency between them, and latent interventions produce visual changes consistent with this direction. These results provide qualitative support that the learned transition structure captures meaningful directed dependencies rather than only preserving information about the physical state.

J DETAILS OF VISUAL TRANSFER EXPERIMENTS

This section provides additional details of the held-out robot transfer experiment on RoboNet (Dasari et al., 2019). The experiment evaluates whether action-conditioned latent dynamics learned from a set of robot platforms remain predictive on robots not observed during training.

Dataset and split. RoboNet contains image-based manipulation data collected from multiple robot platforms. We train all models on Baxter, Kuka, and WidowX and evaluate them without adaptation on held-out Franka and Sawyer robots. Images are resized to 224×224, and all selected subsets use a four-dimensional action input. The source training set contains $9{,}300$ transition samples. We restrict the experiment to robot subsets with compatible action interfaces so that transfer can be evaluated without introducing additional action mappings.

Models and training. For the RoboNet transfer experiments, A-JEPA uses the same compact CNN encoder as in the visual benchmarks, with latent dimension 8. The encoder contains four convolutional blocks with channel widths 32, 64, 128, and 256. Each block applies a 5×5 convolution with stride 2 and padding 2, followed by BatchNorm and $\mathrm{LeakyReLU}(0.1)$. The resulting feature map is globally average pooled and projected to the latent space through

$$\mathrm{Linear}(256, 512) \to \mathrm{LeakyReLU}(0.1) \to \mathrm{Linear}(512, 8).$$

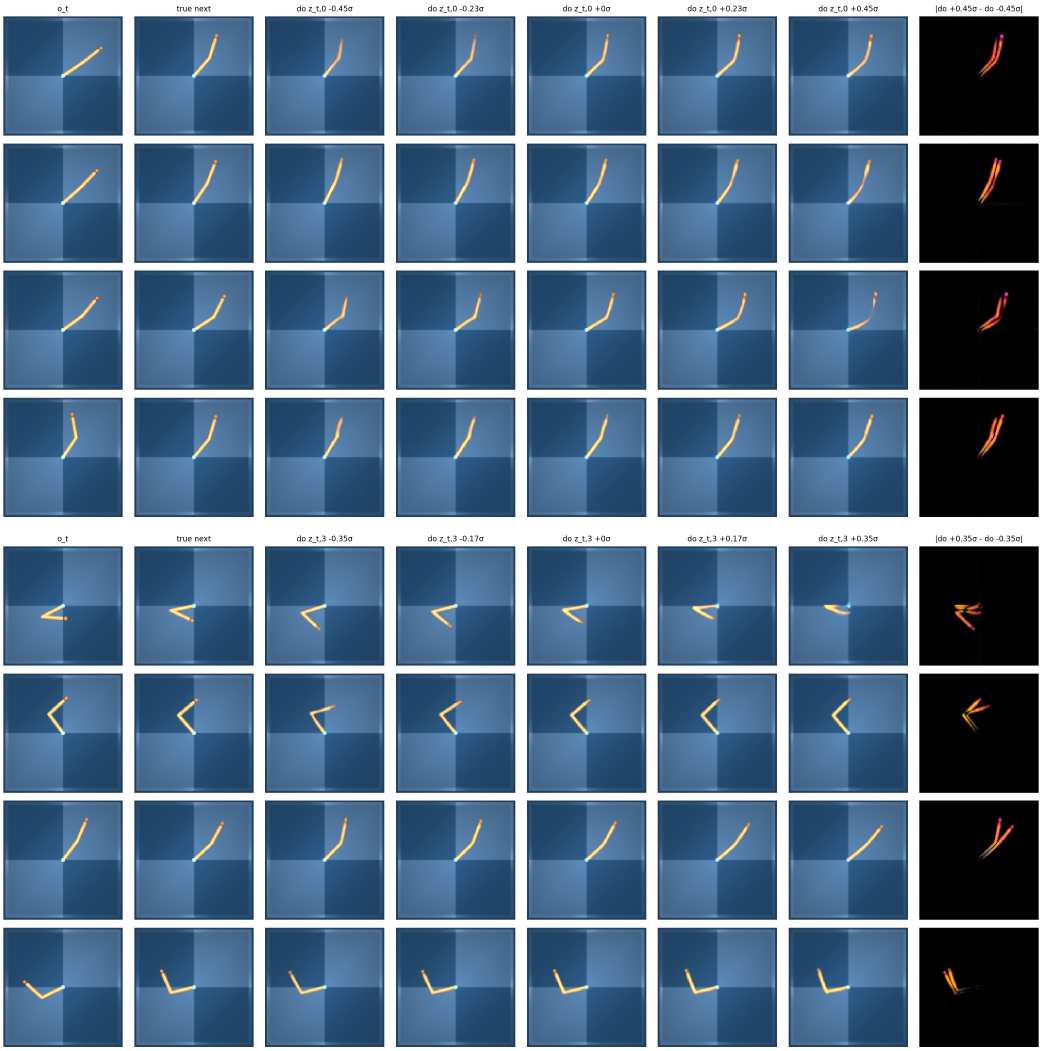


Figure 11: Latent intervention diagnostics on REACHER. Top: intervention on the proximal-aligned coordinate z_0 changes the proximal joint and induces corresponding changes in the downstream distal arm segment. Bottom: intervention on the distal-aligned coordinate z_3 produces a more localized distal-joint change while leaving the proximal configuration relatively unchanged. This asymmetric response is consistent with the learned directed dependency $z_0 \to z_3$.

The four-dimensional action input is mapped to the latent dimension using the same action encoder as in the visual benchmark experiments. A-JEPA uses a coordinate-wise transition model following the implementation in Sec. F. For coordinate i, the conditional mean is predicted from the ordered latent coordinates up to i together with the encoded action, $\mu_{t+1,i} = f_i(\mathbf{z}_{t,\leq i}, \mathbf{a}_t^z)$, where each f_i is a two-layer MLP with hidden width 256. A separate action-conditioned head predicts one log-variance per latent coordinate for the factorized Gaussian transition model. All models are trained for 5 epochs using AdamW with learning rate 10^{-4}, weight decay 10^{-5}, batch size 64, and history size 1. For A-JEPA, we use $\alpha = 0.51$ and $\tau = 1.0$. Results are averaged over three random seeds.

Baselines. We compare A-JEPA with matched-capacity LeWM and PLDM baselines. All methods use the same CNN encoder, latent dimension 8, action input, history size, optimizer, batch size, and number of training epochs. LeWM and PLDM retain their respective predictive objectives and non-DAG transition predictors; they are not given the coordinate-wise transition structure or the Gaussian-contrastive A-JEPA objective.

Table 1: Held-out robot transfer on RoboNet. Models are trained on Baxter, Kuka, and WidowX and evaluated without adaptation on Franka and Sawyer. We report latent rollout R^2 at selected horizons and the average over $h = 1, \ldots, 20$. Values are mean $\pm$ standard deviation over three seeds.

Horizon	A-JEPA	LeWM	PLDM
	Held-out Franka		
$h = 1$	0.9691 ± 0.0088	0.6154 ± 0.2418	0.4486 ± 0.4487
$h = 2$	0.9605 ± 0.0112	0.6033 ± 0.1973	-0.0679 ± 0.4816
$h = 3$	0.9534 ± 0.0128	0.5728 ± 0.1905	-0.8018 ± 1.1256
$h = 4$	0.9468 ± 0.0141	0.5428 ± 0.1928	-1.7901 ± 2.0465
$h = 6$	0.9387 ± 0.0158	0.4878 ± 0.1995	-3.9602 ± 3.8622
$h = 7$	0.9353 ± 0.0165	0.4588 ± 0.2028	-4.8820 ± 4.5269
$h = 8$	0.9314 ± 0.0176	0.4261 ± 0.2080	-5.6401 ± 5.0212
$h = 9$	0.9282 ± 0.0181	0.3948 ± 0.2155	-6.2633 ± 5.3929
$h = 11$	0.9247 ± 0.0192	0.3278 ± 0.2234	-7.2379 ± 5.9311
$h = 12$	0.9227 ± 0.0187	0.2931 ± 0.2235	-7.6042 ± 6.1337
$h = 13$	0.9210 ± 0.0186	0.2582 ± 0.2223	-7.8627 ± 6.2301
$h = 14$	0.9192 ± 0.0191	0.2197 ± 0.2193	-8.0768 ± 6.3313
$h = 15$	0.9170 ± 0.0191	0.1802 ± 0.2132	-8.1900 ± 6.3764
$h = 16$	0.9146 ± 0.0183	0.1445 ± 0.2059	-8.3055 ± 6.4592
$h = 17$	0.9130 ± 0.0184	0.1087 ± 0.2030	-8.4868 ± 6.6007
$h = 18$	0.9115 ± 0.0176	0.0753 ± 0.2007	-8.5143 ± 6.6106
$h = 19$	0.9113 ± 0.0166	0.0426 ± 0.1919	-8.5835 ± 6.6533
$h = 20$	0.9099 ± 0.0167	0.0119 ± 0.1864	-8.7251 ± 6.7523
	Held-out Sawyer		
$h = 1$	0.9816 ± 0.0051	0.8782 ± 0.0225	0.6821 ± 0.0550
$h = 2$	0.9720 ± 0.0074	0.8258 ± 0.0366	0.1957 ± 0.3168
$h = 3$	0.9656 ± 0.0094	0.7963 ± 0.0415	-0.5956 ± 0.9193
$h = 4$	0.9585 ± 0.0111	0.7695 ± 0.0414	-1.7371 ± 1.7133
$h = 6$	0.9514 ± 0.0148	0.7307 ± 0.0406	-4.5519 ± 3.5895
$h = 7$	0.9484 ± 0.0134	0.7118 ± 0.0404	-5.9537 ± 4.3826
$h = 8$	0.9440 ± 0.0145	0.6940 ± 0.0377	-7.2737 ± 5.0679
$h = 9$	0.9408 ± 0.0141	0.6761 ± 0.0360	-8.4299 ± 5.6460
$h = 11$	0.9328 ± 0.0132	0.6383 ± 0.0286	-10.1638 ± 6.4574
$h = 12$	0.9296 ± 0.0121	0.6236 ± 0.0295	-10.6728 ± 6.6503
$h = 13$	0.9276 ± 0.0138	0.6120 ± 0.0285	-11.2198 ± 6.9406
$h = 14$	0.9242 ± 0.0158	0.5993 ± 0.0270	-11.6801 ± 7.1710
$h = 15$	0.9222 ± 0.0163	0.5854 ± 0.0274	-12.0939 ± 7.4304
$h = 16$	0.9194 ± 0.0147	0.5714 ± 0.0288	-12.3088 ± 7.5915
$h = 17$	0.9175 ± 0.0141	0.5614 ± 0.0277	-12.5487 ± 7.7254
$h = 18$	0.9139 ± 0.0144	0.5458 ± 0.0269	-12.8602 ± 7.9320
$h = 19$	0.9087 ± 0.0147	0.5312 ± 0.0326	-13.0918 ± 8.1273
$h = 20$	0.9028 ± 0.0159	0.5155 ± 0.0351	-13.3202 ± 8.2978

Evaluation protocol. At test time, all model parameters are frozen. Given the encoded state $\mathbf{z}_t$ of a held-out robot and its subsequent action sequence, we recursively apply the learned transition model to obtain $\hat{\mathbf{z}}_{t+h}$. At each horizon h, we compare the predicted latent state with the encoded future observation $\mathbf{z}_{t+h} = \mathbf{h}(\mathbf{x}_{t+h})$ using rollout R^2, where larger values indicate more accurate latent prediction. We report rollout R^2 over horizons $h = 1, \ldots, 20$ and its average across horizons.

Results. The main difference lies in long-horizon stability. All models use matched encoder capacity and latent dimension and are frozen when evaluated on robots not observed during training. As depicted by Table 1, A-JEPA maintains rollout R^2 above 0.90 at horizon 20 on both Franka and Sawyer, whereas LeWM degrades substantially and PLDM becomes unstable over longer horizons. These results indicate that the action-conditioned latent dynamics learned by A-JEPA transfer more robustly across changes in robot platform, morphology, and visual appearance.

K Discussion I: LeWorldModel as a Constrained Instance

It is instructive to compare our general information-theoretic objective with the recent LeWorldModel (LeWM) objective (Maes et al., 2026). LeWM optimizes two terms: a next-representation prediction loss and an isotropic-Gaussian regularizer,

$$\mathcal{L}_{\text{LeWM}} = \underbrace{\|\hat{\mathbf{z}}_{t+1} - \mathbf{z}_{t+1}\|_2^2}_{\text{prediction}} + \lambda_{\text{L}} \underbrace{\text{SIGReg}(\mathbf{Z})}_{\text{anti-collapse}}, \tag{84}$$

where $\hat{\mathbf{z}}_{t+1} = \text{pred}_{\boldsymbol{\phi}}(\mathbf{z}_t, \mathbf{a}_t)$. SIGReg encourages the marginal representation distribution to match an isotropic Gaussian through normality tests over random one-dimensional projections. In comparison, our general objective is

$$\mathcal{L}_{\lambda} = \underbrace{-\mathbb{E}\left[\log p_{\boldsymbol{\phi}}(\mathbf{z}_{t+1} \mid \mathbf{z}_t, \mathbf{a}_t)\right]}_{\text{transition predictiveness}} - \lambda \underbrace{H(\mathbf{z}_{t+1})}_{\text{information preservation}}. \tag{85}$$

The two terms in Eq. (84) can be related directly to the two principles in Eq. (85).

Prediction term. Consider a restricted transition model with fixed isotropic covariance,

$$p_{\boldsymbol{\phi}}(\mathbf{z}_{t+1} \mid \mathbf{z}_t, \mathbf{a}_t) = \mathcal{N}\left(\mathbf{z}_{t+1}; \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{z}_t, \mathbf{a}_t), \sigma^2 \mathbf{I}\right). \tag{86}$$

Its negative log-likelihood is

$$-\log p_{\boldsymbol{\phi}}(\mathbf{z}_{t+1} \mid \mathbf{z}_t, \mathbf{a}_t) = \frac{1}{2\sigma^2} \|\mathbf{z}_{t+1} - \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{z}_t, \mathbf{a}_t)\|_2^2 + C, \tag{87}$$

where C is independent of the predicted mean. Hence, with fixed isotropic variance, conditional likelihood maximization reduces exactly, up to scaling and an additive constant, to the MSE prediction objective used by LeWM.

Information-preservation term. LeWM addresses representation collapse by encouraging $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ through SIGReg. This can also be connected to the entropy term in our general objective. In particular, among distributions satisfying

$$\mathbb{E}[\mathbf{z}] = \mathbf{0}, \qquad \text{Cov}(\mathbf{z}) = \mathbf{I}, \tag{88}$$

the isotropic Gaussian is the unique maximum-entropy distribution,

$$\mathcal{N}(\mathbf{0}, \mathbf{I}) = \arg\max_{p_{\mathbf{z}}} H(\mathbf{z}). \tag{89}$$

Therefore, Gaussian regularization can be viewed as a constrained realization of the information-preservation principle in Eq. (85): rather than maximizing entropy over a general admissible representation family, LeWM explicitly drives the representation toward the maximum-entropy distribution associated with fixed zero mean and identity covariance.

Taken together, these observations show that LeWM can be viewed as a constrained instance of our general information-theoretic formulation: its MSE term corresponds to conditional likelihood maximization under a fixed-isotropic Gaussian transition model, while SIGReg realizes entropy-based information preservation under prescribed first- and second-order moments. This interpretation concerns the underlying objective principle; SIGReg is not algebraically identical to the entropy term itself.

What is missing for component-wise identification? This connection also clarifies why strong predictive performance and non-collapsed representations do not by themselves imply recovery of individual causal states. Under an orthogonal transformation $\mathbf{R}$,

$$\|\mathbf{R}\hat{\mathbf{z}} - \mathbf{R}\mathbf{z}\|_2^2 = \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2, \tag{90}$$

and an isotropic Gaussian is likewise invariant under orthogonal rotations. Thus, the LeWM objective does not, by itself, distinguish a latent representation from its orthogonal mixtures. Our identifiability analysis shows that the additional ingredient needed to resolve this ambiguity is sufficient action-induced variation in the conditional transition mechanisms. A-JEPA therefore extends this predictive and information-preserving view by allowing actions to modulate both conditional means and variances, and uses their variation as an identifying signal that reduces the ambiguity to permutation and component-wise invertible transformations.

L Discussion II: Two Levels of Representation Collapse

Representation collapse is a central issue in joint-embedding predictive learning. We argue that it is useful to distinguish two conceptually different levels of representation failure. The first is the conventional *trivial collapse* widely studied in the JEPA and joint-embedding literature, where the learned representation discards information about the underlying state. The second is a more subtle failure that can remain even after such information loss has been prevented: the representation may preserve the underlying state information globally, while mixing multiple latent factors within individual representation coordinates. We refer to this second level more precisely as *representation mixing*.

Level I: information-loss collapse. The conventional representation-collapse problem arises when distinct observations are mapped to identical or nearly identical representations. In the extreme case, $h(\mathbf{x}) = \mathbf{c}$ for all $\mathbf{x}$, such that the predictive objective can be satisfied by a trivial solution without retaining meaningful information about the underlying state. For an action-conditioned predictive model, for example, if $\mathbf{z}_t = \mathbf{z}_{t+1} = \mathbf{c}$ for all observations, a predictor that always outputs $\mathbf{c}$ can achieve perfect representation-space prediction while learning no useful state representation. This is fundamentally an information-loss problem: the variation that distinguishes different underlying states has been discarded by the encoder.

Existing joint-embedding methods address this first level through different anti-collapse mechanisms. I-JEPA uses an asymmetric target-encoder architecture to avoid trivial constant solutions, while VICReg explicitly maintains representation variability and reduces redundancy through variance and covariance regularization (Assran et al., 2025; Bardes et al., 2021). More recently, LeWorldModel (LeWM) uses SIGReg to encourage the marginal representation distribution to approach an isotropic Gaussian, thereby preventing the representation from degenerating to a constant solution (Maes et al., 2026). Although these approaches differ in their implementation, they share the goal of preserving sufficient variation in the learned representation.

Our general information-theoretic objective expresses this requirement directly through marginal entropy maximization:

$$\mathcal{L}_\lambda = -\mathbb{E}\left[\log p_{\boldsymbol{\phi}}(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_{t-1})\right] - \lambda H(\mathbf{z}_t). \tag{91}$$

The entropy term directly opposes concentration of the representation distribution toward a degenerate solution. In particular, under the bounded representation space considered in Theorem 2.1, every global optimum with $\lambda > 1$ attains the maximum possible entropy, with $H(\mathbf{z}_t) = 0$ and $\mathbf{z}_t \sim \mathrm{Unif}((0,1)^d)$. Hence, the constant-representation solution is explicitly excluded at the global optimum. Rather than prescribing a particular architectural or distributional anti-collapse mechanism, our formulation captures this requirement through a general information-preservation principle.

LeWM provides a particularly direct connection to this interpretation. SIGReg encourages $\mathbf{z}$ to follow $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Under fixed first- and second-order moments, $\mathbb{E}[\mathbf{z}] = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{z}) = \mathbf{I}$, the isotropic Gaussian is the unique maximum-entropy distribution. Therefore, Gaussian distribution matching in LeWM can be viewed as a constrained realization of the same entropy-maximization principle. Our practical A-JEPA objective instead implements this principle through the negative-sample contrastive term in Eq. (15), whose connection to representation entropy is established in Sec. E.

Level II: information-preserving representation mixing. Preventing information-loss collapse does not, however, determine how the retained information is organized across representation coordinates. Consider a representation $\mathbf{z} = T(\mathbf{s})$, where T is an invertible transformation of the underlying latent state $\mathbf{s}$. Since T is invertible, no state information need be lost: $\mathbf{s}$ can in principle be fully recovered from $\mathbf{z}$. Nevertheless, each learned coordinate may depend on multiple latent state variables, such that $z_i = T_i(s_1, \ldots, s_d)$. The representation is therefore non-collapsed in the conventional information-preservation sense, while the underlying latent factors remain mixed across representation coordinates.

Importantly, entropy maximization alone cannot resolve this second ambiguity. An invertible transformation can preserve the information contained in the latent state while substantially mixing its individual components. More directly, the change-of-variables formula gives

$$H(T(\mathbf{s})) = H(\mathbf{s}) + \mathbb{E}\log\left|\det J_T(\mathbf{s})\right|. \tag{92}$$

For a volume-preserving transformation satisfying $|\det J_T(\mathbf{s})| = 1$, this reduces to $H(T(\mathbf{s})) = H(\mathbf{s})$. Hence, a fully mixed representation can have exactly the same entropy as the original latent state.

A particularly simple example arises for isotropic Gaussian representations. If $\mathbf{s} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{z} = \mathbf{R}\mathbf{s}$ for an orthogonal matrix $\mathbf{R}$, then $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and therefore has exactly the same entropy and marginal distribution as $\mathbf{s}$. Nevertheless, a general orthogonal matrix $\mathbf{R}$ can mix every learned coordinate across multiple underlying latent variables. Thus, neither information preservation nor maximum entropy alone distinguishes the underlying causal coordinates from their information-preserving mixtures.

Resolving this second level therefore requires additional identifying structure rather than stronger anti-collapse regularization. In our framework, this structure is provided by action-induced variation in the transition mechanisms. As established in Theorem 3.3, sufficiently rich action-induced variation rules out arbitrary cross-component mixing and restricts the learned representation to permutation and component-wise invertible transformations of the underlying latent causal states. In this sense, the identifiability analysis addresses a fundamentally different problem from conventional anti-collapse mechanisms: it determines how the preserved state information is organized across individual representation coordinates.

The two levels therefore require different solutions. The first asks whether the representation retains sufficient information to distinguish different underlying states and can be addressed through anti-collapse mechanisms such as architectural asymmetry, variance preservation, distribution matching, or entropy maximization. The second asks whether the retained information is organized according to the underlying latent factors rather than arbitrary mixtures and requires additional identifiability conditions. In our formulation, these roles are separated explicitly: *entropy maximization prevents information-loss collapse, whereas action-induced mechanism variation resolves information-preserving representation mixing*.